Evaluator Characteristics and Programme Evaluability Decisions: An exploratory study of evaluation practice in South Africa, Brazil, the United Kingdom, and the United States of America

Adiilah Boodhoo

A thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy

Section of Organizational Psychology

UNIVERSITY OF CAPE TOWN

November 2016

Supervisor: Prof Joha Louw-Potgieter

Declaration

I hereby declare that this dissertation is my own work, from conceptualisation to execution. To the best of my knowledge this dissertation contains no material written by another person, except where due acknowledgement has been made. I have not previously submitted this dissertation in its entirety or in part to this university or to any other university.

Signed by candidate

Signature removed

Adiilah Boodhoo

Nov 30, 2016

Date

Acknowledgments

I would like to convey my thanks and gratitude to:

My supervisor, Professor Joha Louw-Potgieter. Thank you for being such an exceptional mentor throughout this very long journey. If it were not for you, I would have taken another five years to complete this PhD. Thank you for keeping me focused and for providing me with so many opportunities to grow professionally and personally. It was an absolute privilege to learn from you. Thanks to you, I walk out of this PhD more empowered and ambitious than before.

The American Evaluation Association, The UK Evaluation Society, The South African Monitoring and Evaluation Association, and the Brazilian Monitoring and Evaluation Network for giving me access to their members.

Veronica Pais and Marita Lindeque for translating my study instruments into Portuguese.

Katya Mauff and Alexa Heekes for helping me make sense of my data and for introducing me to statistical analyses that I knew very little about at the beginning of this journey. Alexa, I cannot thank you enough for being so selflessly invested in my success and for not letting me get “lost in the deep blue ocean”.

May Ka Wai Ho and Adarsh Beeharry for helping me compile this very temperamental document. Adarsh, while I do not tell you nearly enough, I am truly grateful for your unwavering support and faith in me. Thank you for helping me make yet another dream come true.

My dearest friends Ridwana Timol, Aashika Rana, Carren Field, Ines Meyer and Aditi Hunma, and my aunt Nawsheen Elaheebocus. Thank you for walking by my side all the way to the finish line. Each of you have inspired me to achieve some form of greatness and taught me the true meaning of resilience. Aunty Nawsheen thank you for leading the way and showing by example that I too “can steer myself any direction I choose”.

My colleagues and friends in the Section of Organizational Psychology at UCT, and my peers at Claremont Graduate University. You have been an invaluable source of support throughout this journey. Thank you for sharing your own PhD experiences with me – they helped me make sense of mine and inspired me to “keep calm and keep writing”. A special thanks to my Head of Section, Associate Professor Suki Goodman, for accommodating me when I needed time and space to prioritise my PhD.

Mum and dad, Aunty Goree and Uncle Rajesh. There are simply not enough words for me to express my gratitude to you. Let me start by saying: THIS ONE IS FOR YOU. Thank you for all the sacrifices that you have made so that I could pursue my ambition to study abroad and live the life that I want. You instilled in me the belief that I could achieve anything I set my mind and heart to. Thank you for understanding how important this PhD was and still is to me. Mum and dad, I would have never reached this far without your unconditional love and blessings.

Husain and Yusuf, your dry humour and tough love kept me sane and drove me crazy at the same time. Thank you for all the brotherly shenanigans. They were “refreshing” during stressful times.

Abbreviations

AC	Audit Commission
AEA	American Evaluation Association
AFREA	African Evaluation Association
BMEN	Brazilian M&E network
BRICS	Brazil, Russia, India, China and South Africa
CA	Correspondence Analysis
CDC	Centre for Disease Control and Prevention
CPRS	Central Policy Review Staff Cabinet
CSR	Comprehensive Spending Review
DFID	UK Department for International Development
DPME	Department for Planning, Monitoring and Evaluation
DSO	Departmental Strategic Objectives
DWP	Department of Work and Pensions
EA	Evaluability Assessment
EBRD	European Bank for Reconstruction and Development
EN	Evaluation Network
ERS	Evaluation Research Society
FLM	Falsifiable Logic Model
GAO	U.S. General Accounting Office
GDP	Gross Domestic Product
GNI	Gross National Income
GPRA	Government Performance and Results Act
GSRS	Government Social Research Service
GWM&ES	Government-wide Monitoring and Evaluation System
HDI	Human Development Index
ICAI	Independent Commission for Aid Impact
ILO	International Labour Organization
IMF	International Monetary Fund
INTRAC	International NGO Training and Research Centre

IOCE	International Organization for cooperation in Evaluation
JCSEE	Joint Committee on Standards for Educational Evaluation
JMU	Joint Management Unit
LAC	Latin American and Caribbean Countries
LGMA	Local Government Modernisation Agenda
LPSA	Local Public Service Agreement
M&E	Monitoring and Evaluation
MLR	Multinomial Logistic Regression
NAO	National Audit Office
NDM	Naturalistic Decision Making
NEET	Neither Employed nor in Education or Training
NEPF	National Evaluation Policy Framework
NGO	Non-governmental Organization
NPM	New Public Management
NPR	National Performance Review
OECD	Organization for Economic Co-operation and Development
OMB	Office of Management and Budget
PAR	Programme Analysis and Review
PART	Program Assessment Rating Tool
PCA	Principle Component Analysis
PEMD	Programme Evaluation and Methodology Division
PES	Public Expenditure Survey System
PPA	Pluri-Annual Plan
PSA	Public Service Agreement
RFP	Requests for Proposals

SA	South Africa
SAMEA	South African Monitoring and Evaluation Association
SPSS	Statistical Package for the Social Sciences
SR	Spending Review
UK	United Kingdom
UKES	UK Evaluation Society
UNDP	United Nations Development Programme
UNIFEM	United Nations Development Fund for Women
USA	United States of America
U.S.	United States
USAID	U.S. Agency for International Development

Table of Contents

List of Figures

List of Tables

ABSTRACT

Responding to recent calls in the literature for cross-country comparisons of evaluation practice, this simulation study investigated (a) evaluators' perspectives on what determines a programme's evaluability, (b) what criteria evaluators prioritise when assessing a programme's evaluability, and (c) the degree to which practice context (developing, developed, or both) and self-reported levels of evaluation experience predict programme evaluability decisions. Valid responses from evaluators practising in the United States of America (*n* = 94), the United Kingdom (*n* = 30), Brazil (*n* =91) and South Africa (*n* = 45) were analysed. Q factor analyses using data collected via a Q Sort task revealed four empirically distinct evaluability perspectives. The dominant perspectives were labelled as *theory-driven* and *utilisation-focused*. Correspondence analyses demonstrated that participants used different criteria to assess the evaluability of three fictitious evaluation scenarios. Multinomial regression analyses confirmed that practice context and level of experience did not predict the type of evaluability criterion prioritised in any of the scenarios. Evaluators practising in developed countries were more likely to characterise a programme with robust structural features, unfavourable stakeholder characteristics, and unfavourable logistical conditions as *evaluable with high difficulty* than as *evaluable with medium difficulty*. Evaluators with limited experience were more likely than unlikely to embark on an evaluation of such a programme. This study represents the first empirical investigation of how evaluators from selected developed and developing countries assess programme evaluability.

Keywords: evaluability assessments, evaluability criterion, evaluability perspectives, programme evaluability

CHAPTER ONE

INTRODUCTION

Evaluators work within complex, messy, and dynamic environments, each characterised by a unique set of real-world constraints and contingencies. Schwandt (2003, p. 353) refers to these types of environments as the “rough ground”. Given the nature of their work context, evaluators continuously engage in a “complicated juggling act involving trade-offs between available resources and acceptable standards of evaluation practice” (Bamberger, Rugh, & Mabry, 2012, p.7). Such standards of evaluation practice are derived from the knowledge accumulated in the field by both theorists and practitioners, and apply to a myriad of decisions that evaluators have to make. These include: which evaluation questions to prioritise, which methods to use, whom to involve in the evaluation process, and when and how to disseminate the evaluation findings (Miller, 2010).

Different concepts have been used interchangeably in contemporary evaluation literature to articulate theorists’ and practitioners’ notions of how these decisions are to be taken and how evaluation should be practised in general (Donaldson & Lipsey, 2006; Kundin, 2008). These include models of evaluation, theories of evaluation, evaluation, paradigms and evaluation frameworks. The role of such articulations is a contentious matter, with some distinguished evaluators, such as Scriven, dismissing their importance, while others view them as fundamental to our professional entity, and a “central thread in the social fabric of the evaluation profession” (Donaldson & Lipsey, 2006, p. 61).

Evaluation theorists distinguish between prescriptive and descriptive articulations (hereafter referred to as theories). Prescriptive theories consist of a set of prescriptions that implicitly/explicitly specify what evaluators should or should not do as part of their everyday practice for a good evaluation to follow (Alkin, 2004). Descriptive theories, on the other hand, characterise what unfolds in practice and represent different validated possibilities for conducting evaluations. Evaluation theories are largely prescriptive in nature and, more often than not, lack operational specificity and address practice in abstract terms (Christie & Azzam, 2005; Miller,

2010). In other words, they do not articulate explicitly how they can be implemented in practice, and are hence subject to varied interpretations and applications. This abstraction is understandable, to some extent, given the nature of the contexts in which evaluation is practised. Our practice is not amenable to a set algorithm or a template for action. It calls for an improvisational approach as opposed to formulaic one (Schwandt, 2003). According to Greene (2006, p.111), practice decisions are negotiated in a "discretionary space" and are informed by both an evaluator's philosophical adherence to a given theory and the presenting features of the evaluation context. As such, evaluation practice and evaluation theory do not intertwine closely in many instances or overlap at all in others (Chelimsky, 2013).

Evaluation practitioners tend to view the propositions of early theorists as divorced from reality and failing to incorporate the real-world complexities imposed by evaluation contexts. As a result they have developed their own implicit and pragmatic theories to guide their practice decisions. Consequently, it seems important to investigate systematically what "folk theories" exist around different areas of practice, and what guides evaluators' practice decisions (Christie, 2003a, p.92).

Interestingly, the few studies on evaluation practice have identified a consistent pattern: evaluators do not necessarily conform to established evaluation theories in their everyday practice but draw on their experience, training and practical reasoning to make practice decisions (Christie, 2003b; Kundin, 2008; Shadish & Epstein, 1987; Tourmen, 2009). Is this the case across all areas of programme evaluation practice and for all evaluators, irrespective of their practice context?

Empirical investigations of specific evaluation practices are limited and sporadic (Demarteau, 2002; Kundin, 2010), and tend to concentrate on evaluation utilisation and influence (e.g., Altschuld, Yoon, & Cullen, 1993; Boyer & Langbein, 1991; Braskamp, Brown, & Newman, 1982; Christie, 2007), and evaluation design (e.g., Azzam, 2011; Azzam & Szanyi, 2011).

The empirical literature on programme evaluability is sparse and little is known about how evaluators operationalise prescriptive theories of evaluability (Watts &

Washington, 2016). The purpose of the present study is to explore inductively and comparatively how different cohorts of evaluators conceptualise and operationalise programme evaluability, and whether or not their operationalisations are consistent across evaluation contexts, and in line with how they think they assess evaluability. Given that this is a novel investigation into an under-studied and emerging area of evaluation, an inductive approach is preferred over a deductive one.

By manipulating systematically selected evaluability conditions within fictitious evaluation scenarios, it is possible to examine how and if evaluators reshape their operationalisations of programme evaluability depending on the features of the evaluation context. By comparing evaluators who practice in developing and developed contexts, and with varying levels of experience, it is possible to gain insight into how these characteristics affect evaluators' programme evaluability decisions and the way they operationalise evaluability. The present study uses this very approach to address the following research questions:

1. Do evaluators share a common perspective towards evaluability? If not, what perspectives can be empirically identified and what evaluator types are most associated with these perspectives?
2. Are evaluators' prioritisation of evaluability criteria consistent across different study tasks (i.e., three different evaluation scenarios, and one a-contextual sorting task)?
3. Do selected evaluator characteristics (practice context and experience) predict their evaluability assessments, likelihood to evaluate, and prioritisation of evaluability criteria?

Conceptual Framework of Current Study

This study draws on the different elements of a conceptual framework proposed by Kundin (2010) for studying evaluators' practice decisions (see Figure 1). This framework consolidates the work of multiple theorists (e.g., Fournier, 1995; Greene, 2005; Hansen, 2005) and isolates three key elements that might shape an evaluator's decision-making process: situation awareness, practical reasoning, and reflection in action.

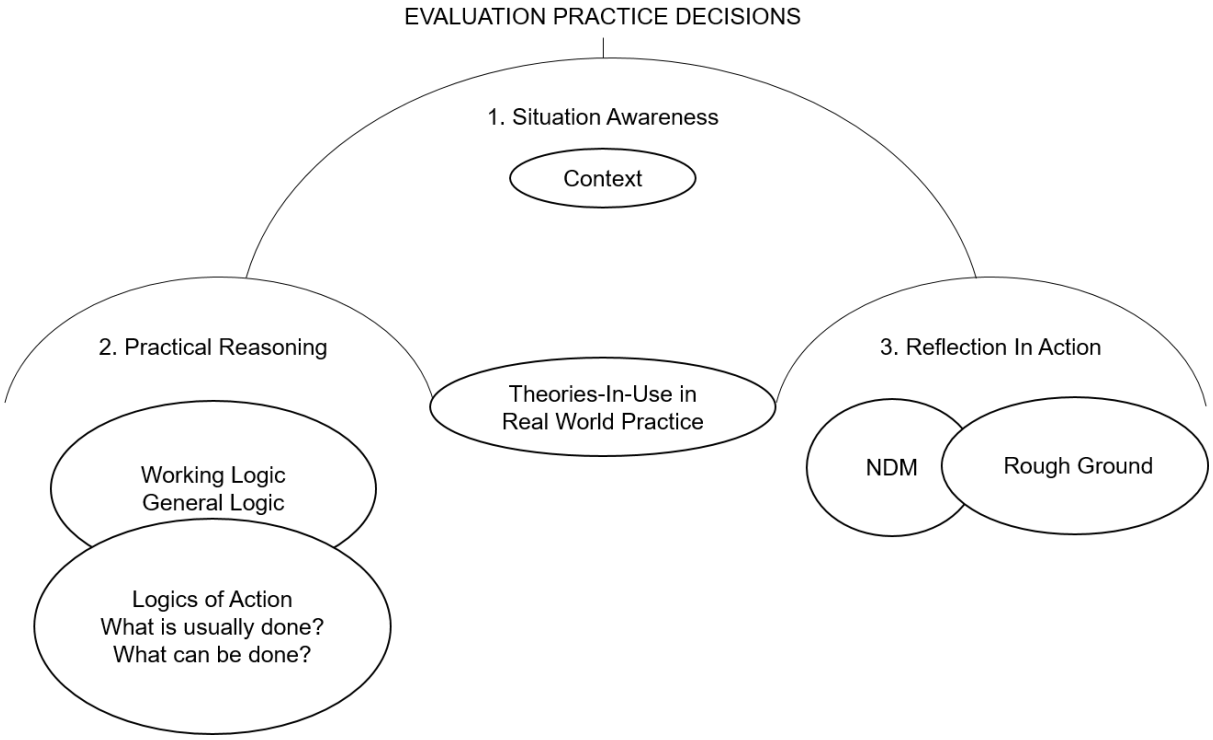


Figure 1. Conceptual framework for studying evaluators' practice decisions

Situational awareness implies an appreciation of the inherent features and constraints of an evaluation context and its associated dimensions, including, the descriptive and economic features of the setting, the institutional and organizational climate, and the related interpersonal and political dynamics. An evaluator's ability to assess an evaluation context accurately would, in principle, facilitate his/her task in choosing between competing alternatives.

When evaluators use practical reasoning (the second element of the framework), they draw on both explicit and tacit procedural knowledge to respond to volatile and complex evaluation situations. Their explicit and tacit procedural knowledge is informed by their academic training and socialisation, their cumulative experience in the field, and their past evaluations (Fournier, 1995; Greene, 2006). Evaluators also consider *what is usually done* and *what can be done*, given a particular set of contingencies. Hansen (2005) refers to this approach as logics of action (a dimension of practical reasoning). Here evaluators draw on their past successes and response repertoire to replicate decisions associated with positive outcomes. Alternatively, they choose the course of action that they feel competent to undertake.

Evaluation decisions are also driven by a general logic and a working logic in evaluation. The first type of logic “specifies the game and the rules of the game that one is playing when conducting an evaluation in any field” (Fournier, 1995, p.17). These include: establishing criteria of merit, constructing standards, measuring performance against standards, and making a judgement of merit or worth. The second type of logic, also referred to as logic in use or reconstructed logic, represent the different operationalisations of general logic. In other words, they represent the variations in application of general logic in practice. For example, evaluators might use different approaches to identify criteria of merit, construct standards, measure performance, and synthesise data. Their working logic might be informed by a multitude of factors, including their own implicit theories of real world practice.

Reflection in action (Schon, 1983), the third element of the framework, refers to the active and systematic process of reflecting on our practice as it unfolds on the “rough ground” (Schwandt, 2003, p. 353) and after an evaluation has been completed. Reflective practice, identified as a key evaluator competency, enables recognition of the assumptions, theories, and paradigms that underlie our actions and decisions, and allows for continuous learning (Stevahn, King, Ghere, & Minnema, 2005). Kundin (2010) advocates the use of a naturalistic decision making (NDM) framework to encourage evaluators to engage in reflective practice and articulate how they make decisions in real-world settings. NDM-based research takes place in the complex environment of the decision maker (as opposed to simulated and controlled settings) and involves real-time field observations of how consequential decisions are taken (Lipshitz, Klein, Orasanu, and Salas, 2001).

This study used a line of inquiry that incorporates most aspects of Kundin’s (2010) conceptual framework to investigate how evaluators make evaluability decisions. The study methods included three fictitious evaluation scenarios that mimicked the nature of real world evaluation contexts, with each scenario embodying both favourable and unfavourable conditions. This approach allowed evaluators to engage in situated decision-making. While the study did not use an NDM approach, it required evaluators to engage in reflective practice in order to explain their decision for characterising a particular scenario as *evaluable with minimal difficulty* or

evaluable *with a lot of difficulty*. Finally, evaluators had to display their working logic by prioritising a set of evaluability criteria in a Q Sort task. The evaluability criteria presented to evaluators in the Q Sort task were derived from the literature and can thus be conceptualised as the general logic of evaluability.

Significance of Study

We know very little about how evaluation is practised on the rough ground (Christie, 2011; Henry & Mark, 2003; Miller, 2010). Descriptions of what evaluators actually do in practice are not explicitly articulated in evaluation reports. In fact, "most technical documents read as smooth, polished, and error-free […] [and do not reflect] the logic in use that would allow us to see many of the most important incidents in the drama of science, which reconstructed logic leaves safely behind the scene" (Worthen, 1995, p. 166). Although calls for empirical investigations of evaluation practice "have struck a chord in some quarters of the evaluation community, these calls have been infrequently answered" (Miller, 2010, pp.390), until recently (Coryn et al., 2016). There is a growing and longstanding concern that prescriptive theories of evaluation are not empirically grounded and contextually relevant (Christie, 2011; Smith, 1993). A systematic understanding of *what* evaluators do in practice and *why* can inform the development of: (a) contingency theories that specify the conditions under which specific evaluation practices would work best, and (b) descriptive theories that provide validated practice possibilities to evaluators for improving their evaluations (Christie, 2011; Mark & Henry, 2003). The relevance of this type of investigation is well-recognised by both practitioners and theorists, as evidenced in a recent study conducted by Coryn et al. (2016). The overwhelming majority of sampled evaluators and prominent theorists (affiliated to the American Evaluation Association; AEA) indicated that research on evaluation has influenced their thinking about the discipline/profession and their own evaluation practice.

The present study is an empirical investigation of evaluation practice. To break away from the tradition of accumulating research on evaluation use (Johnson et al., 2009, located 41 empirical studies on evaluation use conducted between 1986 and 2005), this study focuses on programme evaluability, a concept that emerged in the literature around the same time as evaluation use, but which is arguably less

investigated. There has been, however, a renewed interest in the concept of evaluability and its application, more specifically in the form of evaluability assessments, EAs (Davies, 2013; Trevisan, 2007). This renewed interest is evidenced by recent EA-related publications, such as peer-reviewed articles (e.g., Walser & Trevisan, 2016), and textbooks (e.g., Trevisan & Walser, 2014). Evaluability has also been a frequent subject of discussion on online evaluation blogs and forums, such as EVALTALK and EA365 (sponsored by the American Evaluation Association), and MandE News (hosted by Rick Davies), in recent years. Figure 2 depicts the trend in EA-related publications per year, over the period 1979 to 2012 (Davies & Payne, 2015).

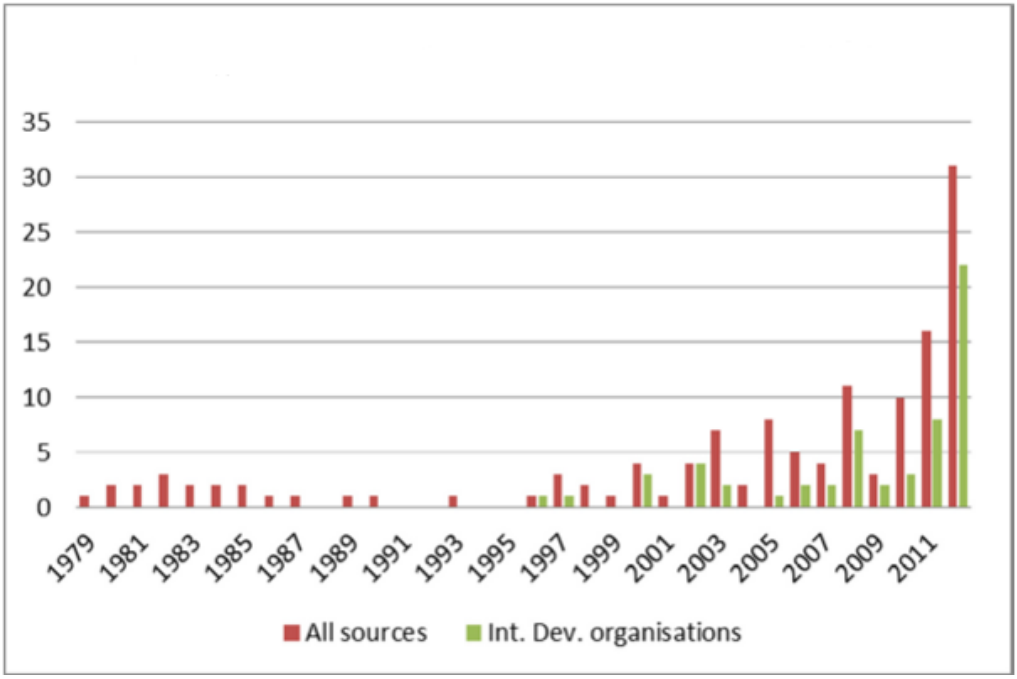


Figure 2. Trends in publications of evaluability assessments.

The present study addresses a domain of inquiry that is particularly relevant from a pragmatic and ethical standpoint. Evaluations are not only resource intensive but have an inherent political and ethical dimension (Schwandt, 2007). Malbry (1997, as cited in Smith, 1998, p.180) refers to evaluation practice as "the most ethically challenging [approach] to research inquiry", and cautions evaluators not to rush into conducting an evaluation. While any programme can be evaluated in some way, at some level, and at a certain cost (Finckenauer, Margaryan, & Sullivan, 2005), it is important for evaluators to deliberate on the following: Is it ethical or pragmatic to

commit evaluation resources to programmes that do not meet the minimum criteria for evaluability? Is it ethical to accept an evaluation contract that one might not be able to honour because of limited evaluation resources? Is it professionally viable to conduct an evaluation that does not comply with accepted standards of practice?

According to Smith (1998), an evaluation contract should be declined if it is not feasible to conduct the evaluation at an acceptable level of technical quality. This is in line with the American Evaluation Association's (2004) first guiding principle for evaluators. It can be argued that the likelihood of violating professionally accepted standards of practice is higher when evaluating programmes with low evaluability. What evidenced-based criteria can evaluators therefore use to identify these programmes? This is the one of the questions that this study seeks to address.

The ability to assess programme evaluability has been identified as a key evaluator competency (under the situational analysis domain) in Stevahn et al.'s (2005) *Essential Competencies for Program Evaluation* taxonomy. This competency has also been officially endorsed by the Canadian Evaluation Society (under the technical practice domain) and the Aotearoa New Zealand Evaluation Association (under the contextual analysis and engagement domain). While the present study does not investigate the extent to which participating evaluators possess this competency, it explores whether or not they have a common or divergent understanding of evaluability. Patterns identified might spur recommendations for evaluator training programmes.

This study represents to my knowledge the first empirical study of how evaluators from four countries of interest assess programme evaluability. It is a direct response to Henry and Mark's (2003) call for more comparative research on evaluation practice. The study is also designed to isolate evaluator characteristics that influence programme evaluability decisions. The purpose is not to judge the accuracy of these decisions. Although simulation studies have been used to examine evaluation practice (e.g., Azzam & Szanyi, 2011; Azzam, 2011; Alkin & Christie, 2005), few have collected data from evaluators outside of the United States, and attempted to categorise these responses in order to identify similarities and differences in evaluation practice. This study presents a distinct opportunity to contrast empirically

four very different cohorts of evaluators and add to the limited body of empirical knowledge on programme evaluability in four countries of interest.

CHAPTER TWO

PROGRAMME EVALUABILITY

Many evaluations arrive at no-effect/inconclusive findings, or culminate into reports that, at best, provide a mere account of programme deficiencies (Leviton, Khan, Rog, Dawkins &, Cotton, 2010; Smith, 1990). This often occurs when evaluators work with programmes that might warrant an evaluation but do not necessarily meet the minimum criteria or standards for evaluability (Wholey, 1979). For ease of reference, these programmes will be referred to as unevaluable programmes. This chapter first provides a brief account of how the concept of evaluability emerged in the literature, followed by an explanation of the concept, its significance, and how it relates to concepts of usefulness and evaluation use. The key differences between these two concepts are delineated. The literature on evaluability assessments is then presented. This chapter culminates into the development of an evaluability framework that captures the key evaluability criteria derived from the literature. Each variable within this framework is then operationalised.

Emergence of the Concept and Key Definitions

The concept of programme evaluability emerged in the literature in the 1970s, as part of the early debates on the questionable utility of evaluations in informing budgetary and policy decisions, poor utilisation of evaluation findings, misuse of evaluation findings, and the difficulty in evaluating poorly conceived programmes (Horst, Nay, Scanlon, & Wholey, 1974; Rutman, 1980; Schmidt, Scalon, & Bell, 1979; Strosberg & Wholey, 1983; Trevisan, 2007; Trevisan & Huang, 2003; Wholey, Nay, Scanlon, & Schmidt, 1975). Rutman (1980) argued that many of these concerns, including the questions raised about the usefulness of programme evaluation, could be attributed to the lack of methodological rigor in many evaluation studies. To illustrate this point, Rutman (1980) drew on Bernstein and Freeman's (1975) assessment of federally funded evaluations. These authors concluded that about half of the studies that measure impact were deficient either in design, sampling, or validity. In addition to methodological concerns, there are other major obstacles that inhibit the utilisation of evaluation findings (even if a study is

technically rigorous, relevant, timely, and properly communicated to decision makers). For instance, there might be strong public acceptance of some programmes, regardless of their value or programme stakeholders might be predisposed to maintain the status quo, irrespective of evaluation findings (Rutman, 1980).

The issues discussed above relate to the concept of programme evaluability and the common practice of conducting premature evaluations. In the present study, evaluability is conceptualised as *the extent to which a programme is ready for useful evaluation*, as opposed to t*he extent to which it can be evaluated in a reliable and credible manner*. While the second definition is widely used by international development agencies (Davies, 2013), it has a major shortcoming. It is argued in the literature that any programme can be evaluated in some way, at some level, and at a certain cost, but not all programmes are ready for useful evaluations (Finckenauer et al., 2005; Schmidt et al., 1979; Wholey, 2010). When determining the evaluability of a programme the question is therefore not whether a programme can be evaluated but whether a programme is ready for useful evaluation. According to Newcomer, Hatry, and Wholey (2010) five basic questions should be asked when any programme is being considered for evaluation:

- Can the results of the evaluation influence decisions about the programme?
- Can the evaluation be done in time to be useful?
- Is the programme significant enough to merit evaluation?
- Is the programme performance viewed as problematic?
- Where is the programme in its development?

In terms of the first criterion, programmes for which decisions must be taken about their continuation, modification, or termination, are better suited for an evaluation than programmes that have considerable political support. In terms of the second criterion, if an evaluation cannot be completed in time to affect programme decisions, the evaluation will not be useful. In terms of the third and fourth criteria, programmes that are resource intensive and/or whose performance is perceived as problematic are good candidates for evaluation, assuming that criterion one and two have been

met. New/pilot programmes for which costs and benefits are unknown are also good candidates for useful evaluation (Newcomer et al., 2010).

Grasso (2003), on the other hand, argues that evaluators have to address the following three questions to make an evaluation useful:

- Who will use the evaluation?
- What will they need from the evaluation?
- When will they need the information?

Answering the first question requires identifying, at the beginning of the evaluation process, who is likely to use the evaluation findings. For an evaluation to be useful, there must be a demand for it (Feinstein, 2002). Identifying the target audience is the first step in gauging that demand. The next step involves aligning the information needs of the various audiences (demand) with what the evaluator can actually deliver (supply). This recommendation is in line with Rutman's (1980) assertion that for an evaluation to be useful, the need for evaluative information must be well established, and the evaluation must produce information relevant to decision makers. Prioritising potentially conflicting information needs, and soliciting consensus and buy-in from the key stakeholders are therefore critical tasks of evaluators (Grasso, 2003). The third question addresses the issue of conducting evaluations that capture the information needs of key stakeholders, but are not well-timed. For an evaluation to be useful, it has to be completed within a timeframe that ensures maximum impact. In addition, the scope and level of evaluation must be feasible within this specified timeframe (Strosberg & Wholey, 1983; Wholey, 2010).

Key Difference between Concepts of Evaluability, Usefulness, and Evaluation Use

Conceptual overlaps between evaluability, usefulness and evaluation use are evident. Authors such as Schmidt et al. (1979), Finckenauer et al. (2005) and Wholey (2010) define evaluability in terms of the extent to which a programme is ready for useful evaluation. Key considerations that evaluators must take into

account to enhance the usefulness of evaluations have been discussed. These include: identifying the target audience for the evaluation, generating information that is relevant to the target audience, and ensuring this information is available when needed. None of these considerations, however, guarantee that the results of the evaluation will be used (Grasso, 2003). Evaluation use is a multifaceted construct that encompasses a number of dimensions, namely instrumental, conceptual, and symbolic (Caracelli, 2000; Shulha & Cousins, 1997; Weiss, 1998). Instrumental use relates to the use of evaluation findings to guide decision-making. Instrumental use is particularly common under the following circumstances: (a) if the implications of the findings are relatively non-controversial; (b) if the recommended changes are minimal; and (c) if the programme is relatively stable, with few changes in leadership, budget, or types of beneficiaries/recipients served (Weiss, 1998). Conceptual use relates to the educative function of evaluations, whereby stakeholders who engage with the evaluation process have a more refined understanding of what the programme is and does (Shulha & Cousins, 1997; Weiss, 1998). Symbolic use of evaluations is political in nature. In this particular scenario, evaluations are used to mobilise support and legitimatise the position of key stakeholders in relation to required programme changes.

While the concept of evaluability embodies aspects of usefulness (i.e., does the programme lend to useful evaluation?) and evaluation use (i.e., how will the evaluation be used?), the concept also has a number of other dimensions. Davies (2013), for example, argues that the concept has two distinct components (other than utility). The first component, labelled *in principle evaluability*, relates to the nature of the programme design and theory of change. The second component, labelled *in practice evaluability*, relates to the availability of relevant data and data capabilities. It should also be noted that while usefulness and evaluation use can be conceptualised as *necessary conditions for evaluability*, they are not sufficient conditions for evaluability.

Key Impediments to Conducting Useful Evaluations

Horst et al. (1974) identified three key deficiencies that might compromise the evaluability of a programme and the usefulness of evaluation efforts:

- Lack of definition: when the problem to be addressed, the programme intervention, and the desired programme outcomes are not sufficiently defined to be measurable. Examples of vague programme language often used to describe interventions include *integrated services*, *a range of modalities*, and *coordinative mechanisms*.
- Lack of clear logic: when the underlying logic of the programme (i.e., the assumptions that link programme input, programme intervention, programme outcomes, and resulting impact) is not well-understood or specified.
- Lack of management: when those in charge of the programme lack motivation, understanding, ability or authority to facilitate the evaluation process and act on evaluation findings.

If a programme has one or more of these deficiencies, there is a low probability that an evaluation will be useful. For example, if there is a lack of management, even findings from a high quality evaluation are not likely to be used for programme improvement (Horst et al., 1974). If there is a lack of definition, different evaluations of the same programme might not be comparable. In addition, it is difficult to propose a well-defined solution to a problem that is ill-defined, and more difficult to accurately evaluate the success of that proposed solution. The quality, value, and usefulness of an evaluation are therefore dependent on the extent to which these programme deficiencies are addressed.

Other major impediments to conducting useful evaluations include flawed programme design and implementation (Davies, 2013; Kaufman-Levy & Poulin, 2003; Van Voorhis & Brown, 1996). A programme can have a flawed design or result in inconsistent/flawed implementation when:

- The programme is not delivered according to design. This can occur, for instance, when documented descriptions of the programme are not detailed enough to facilitate its consistent implementation or when there is no formal programme design in place.
- The programme has intuitive appeal but no theoretical or empirical underpinnings.
- Programme staff do not understand the programme. This can occur, for instance, if they were not part of the planning process or were not properly trained to implement the programme consistently.

Consequences of Evaluating Unevaluable Programmes

There are two possible consequences of evaluating unevaluable programmes:

- Inconclusive evaluation findings.
- Evaluation findings that do not address the information needs of programme stakeholders. Even if the evaluation is methodologically sound, those in charge of the programme may find it irrelevant to their decision making context.

These consequences tie in with what Scanlon et al. (1974; as cited in Smith, 1989, p.15) referred to as “measuring something that does not exist and measuring something that is of no interest to management and policy makers”. Reporting on the impacts of undefined or unevaluable programmes is synonymous to black box evaluations (Rutman, 1980). The issue here is that, since unevaluable programmes are often not properly described, these evaluations do not provide any basis for replicating successful programmes or avoiding ineffective ones. Also, the existence of poorly defined programmes implies that a distinction cannot be made between a poorly implemented programme and an ineffective one (Rutman, 1980). In addition, many unevaluable programmes are characterised by vague and unrealistic programme goals. Vague goals are often attractive to programme stakeholders for two reasons: (a) they provide them with flexibility to change the programme, and (b) programme stakeholders cannot be held accountable for vague goals as they are

open to varying interpretations. If evaluators redefine these vague goals and use their own discretion in selecting the measures, the evaluation might be flawed on number levels because the programme might have been measured against unrealistic or incorrect criteria (Rutman, 1980).

Horst et al. (1974) provided three recommendations to conclude their discussion on programme evaluability: (a) the evaluator should only evaluate programmes that are evaluable; (b) the evaluator should assist with the definitional problems of potentially evaluable programmes; and (c) the evaluator should advise programme stakeholders on whether their programmes are or are not evaluable, and why.

Evaluability Assessment

There are a number of approaches that evaluators can use to enhance their likelihood of conducting useful evaluations. Many of these approaches explicitly address the need for evaluation planning. Evaluability assessment (EA) is one such approach, both widely discussed in the literature and implemented across a wide variety of programmes, disciplines, and settings (Levition et al., 2010; Trevisan, 2007). EAs cut across the broad range of challenges experienced by evaluators and consumers of evaluations (Rutman, 1980). EA was originally developed by Wholey (1979) as a low cost pre-evaluation activity to assess whether programmes were ready for summative evaluations (Rutman, 1980; Trevisan, 2007; Trevisan & Huang, 2003). Wholey et al. (1975) identified several aspects of federal programmes that compromised the feasibility of such evaluations: (a) poorly defined objectives, (b) insufficient resources to meet programme objectives, (c) ambiguity about what constitutes programme success, (d) no apparent logic that connect programme activities to stated outcomes, and (e) specification of outcomes that cannot be measured. In addition, the complex policy and management environment in which these programmes were developed and implemented created instability in terms of resource allocation, type of information needed, and intended users of evaluations (Jung & Schubert, 1983).

There was however continued pressure from policy makers to produce useful information, even in the absence of well-designed and fully operational programmes.

EA was first introduced as a systematic process to tackle this dilemma and avoid premature investments in impact evaluations. Over the years the purposes of EA have expanded to include: (a) identifying programmes that are not worth evaluating, preventing evaluation attempts that are prematurely terminated due to unanticipated issues; (b) facilitating programme improvement; (c) ensuring that relevant and technically feasible evaluations are conducted; (d) maximising utilisation of evaluation findings; (e) building evaluation capacity; and (f) analysing the feasibility of implementing the desired evaluation design (Leviton et al., 2010; Nay & Kay, 1982; Rutman; 1980; Schmidt et al., 1979; Smith, 1989; Thurston, Graham, & Hatfield, 2003; US Agency, International Development, 2008).

In the last five years, there has been a resurgence in the use of EAs and a substantial increase in the number of EA guidance material published by international donor agencies such as the European Bank for Reconstruction and Development (EBRD), and the United States Agency for International Development (USAID) (Davies, 2013). While there is limited evidence on the effectiveness of EAs, some practitioners argue that even modest improvements to a programme or subsequent evaluations can offset the relatively low costs associated with the EA process.

EA process.

Wholey (1979) conceptualised the EA process as cyclical and iterative. Wholey (1979, 2004) initially enumerated eight steps as part of the EA process, which he later combined into a six-step model. These steps are:

(1) Involving intended users of the evaluation.
(2) Clarifying the intended programme.
(3) Exploring programme reality.
(4) Reaching agreement on any required programme changes.
(5) Exploring alternative evaluation designs.
(6) Agreeing on evaluation priorities and intended uses of information.

The initial step in evaluability assessment is to involve the potential evaluation users, obtain their commitment and define the scope and purpose of the assessment (Leviton et al., 2010).

The assessment team/evaluator then reviews programme documents such as vision and mission statements, written goals and objectives if any, and grant proposals. The assessment team/evaluator concurrently holds interviews with primary stakeholders to clarify the relationships between the programme's resources, activities, and desired outcomes. In the final phase of this step, the assessment team/evaluator uses the information gathered from the document review and interviews to develop a logic model or theory of change that is continually revised and shared with stakeholders as more information is gathered.

Once there is general agreement among stakeholders and evaluators about the logic model, the next step is to investigate the program reality. As part of this step, a comparison is made between the programme design captured in the logic model and the programme's reality. Further interviews, documentation review (including reports of past evaluations), and site visits are conducted for this purpose (Wholey, 2004). If any discrepancies are found between the logic model and programme reality, evaluators usually highlight possible factors that might inhibit effective programme performance and identify realistic and measurable indicators of performance.

The output of an EA is a report that addresses the plausibility of the logic model, areas for further programme improvement, the feasibility of conducting the desired evaluation, and options for evaluation design (including the data to be collected, the costs and timeline associated with such an evaluation (Leviton et al., 2010; Smith 1989). These aspects are discussed with programme stakeholders to assist them to take further decisions regarding the programme. These decisions might range from changing programme design/resources allocated to moving forward with a full-scale evaluation.

The last step ties the EA process together by reaching a shared agreement on evaluation priorities and the intended uses of the information.

Adaptations of EA.

A number of adaptations of Wholey's (1979) original EA model are discussed in the literature and implemented in practice. Twelve different stage models were identified by Davies (2013). Examples of these include Kaufman-Levy and Poulin's (2003) five task model for evaluability assessments, Smith's (1989) ten-step EA model, Thurston and Potvin's (2003) seven-step framework, and a recent evaluability model developed by Trevisan and Walser (2014).

While all adaptations of the EA process are based on applied experiences of evaluation practitioners, definitions and sequence of operations vary (Schmidt et al., 1979). Rutman (1980) argued that paying attention to issues that affect evaluability of programmes is more important than focusing on the mechanics of carrying out a prescribed set of EA steps. He also argued that adaptations of the EA process might be required to suit particular circumstances.

A number of evaluability checklists been developed by practitioners, evaluation units, and international development agencies. These checklists vary in length, content, and structure (Davies, 2013). For instance, the United Nations Development Fund for Women's (UNIFEM, 2009) programme evaluability checklist consists of 17 questions framed around evaluability parameters of programme design, availability of information, and conduciveness of the context, while the checklist developed by Peersman, Guijt, and Pasanen (2015) operationalises three dimensions: plausibility, utility, feasibility of measuring impact.

Some evaluability checklists require evaluators to aggregate individual judgements across different evaluability dimensions into a total score and assess the overall evaluability of a programme accordingly. The International Labour Organization's (ILO) Evaluability Assessment Tool, for example, distinguishes between four levels of evaluability (fully evaluable, mostly evaluable, limited evaluability, not evaluable) based on aggregate weighted scores across six dimensions of evaluability. Peersman, Guijt, and Pasanen's (2015) checklist, on the other hand, allows evaluators to reach one of following broad conclusions: (a) no barriers exist–proceed with impact evaluation, (b) impact evaluation is assumed feasible in the near future–

proceed but address critical issues first, and (c) critical barriers cannot be addressed easily or in a timely manner–do not proceed with impact evaluation.

Limitations of EA as a concept and method.

Authors such as Smith (1990), Trevisan (2007), and more recently Watts and Washington (2016), have critiqued the vague, ambiguous and inconsistent articulation of evaluability as a concept and the lack of clear EA methodology in the literature. While there is continued use of EA across a variety of programmes, disciplines and settings, revisions to the EA process and the additions of new EA models are not well documented and justified (Leviton et al., 2010; Trevisan, 2007). In addition, retrospective reviews of the EA process have revealed a number of inherent obstacles. The EA process is based on an underlying assumption of programme rationality. For example, the process assumes that decision makers can be identified, and that programmes will remain static over the EA process (Smith, 1989, 1990). In reality, these assumptions of rationality do not hold. In addition, evaluators conducting the EA may lose programme objectivity (due to the nature of the process), thus undermining the credibility of subsequent evaluations.

Another limitation of the EA process/method is that it can rarely be applied with the intended systemacy and logic (Schmidt et al., 1979). Evaluators often have less control over evaluation assignments than the process assumes. For example, they might have minimal influence on the development of programme objectives, particularly when these objectives are political responses rather than guides for programme implementation or evaluation. In addition, evaluators often do not choose which programmes to evaluate or set timelines for completion. In many instances, evaluation assignments are based on priorities of policymakers and funding agencies, and are mandated regardless of programme readiness. The mere fact that some programmes might not be evaluable within a given EA framework might not be sufficient to halt or delay certain evaluation efforts (Rasp, 1981; Smith, 1981).

Another recurrent challenge encountered by practitioners is determining what qualifies as enough evidence to conclusively establish that a given evaluability condition has been met (Davies, 2013). For example, it is not clear how much

evidence is needed before assuming that a given context is conducive for an evaluation. It is also not clear which evaluability dimension should be assigned the highest weighting when evaluability checklists are used. This makes any categorical judgement about evaluability difficult (Davies & Payne, 2015).

Common benefits/outcomes of EA.

When properly implemented, EA can however save scarce evaluation resources by recommending evaluation only when programmes are ready, establishing evaluation priorities and targeting evaluation resources for essential programme needs, and providing a front-end look of probable evaluation limitations and obstacles (Rutman, 1980; Trevisan & Huang, 2003). In addition, by formalising the agreements between the evaluator and programme stakeholders about evaluation questions, the programme being assessed, and design to be used, an EA helps protect the credibility of an evaluator (Kaufman-Levy & Poulin, 2003; Levition, 2010).

Common outcomes of EA include: (a) the clarification of programme goals and objectives; (b) the development of a programme theory and performance measures; (c) modification of programme components; and (d) the documentation of stakeholder awareness, understanding, and interest in the programme (Finckenauer et al., 2005; Smith, 1989; Trevisan, 2007; Trevisan & Huang, 2003). The EA process therefore sharpens the focus of a given programme (Leviton et al., 2010). In addition, an EA can assist evaluators to classify a programme under one of three categories: evaluable, potentially evaluable with further programme or management definition, or not evaluable (Horst et. al, 1974). Given the broader scope of EA nowadays, the focus is more on how to get a programme to converge to an evaluable form rather than simply assessing whether a programme is evaluable or not (Nay & Kay, 1982; Trevisan & Walmer, 2014). EAs can serve a formative evaluation purpose, and strengthen subsequent summative evaluations.

Evaluable Programmes

In Wholey's (1974, 2010) view, a satisfactorily completed EA should provide a comprehensive model of an evaluable programme. An evaluable programme has the following characteristics:

- Clearly defined, measurable and agreed upon objectives.
- An explicit and plausible programme theory.
- A set of sequenced activities that are implemented as planned.
- Specified indicators of programme implementation and performance that can be easily measured with available evaluation resources.
- The programme warrants further evaluation based on clearly identified information needs.

Nay and Kay's (1982) model, captured in Table 1, complements Wholey's (1994, 2010) account of what constitute an evaluable programme.

Table 1

Characteristics of Evaluable Programmes

Characteristics	Benchmark
Structure and operational relationships	Defined and in place
Agreed upon key expectations	Plausible and attributable to the direct intervention
Agreed upon potential measurements	Feasible to take
Defined and agreed upon potential comparisons	Feasible to make
Intended users of evaluation results	Capable of acting or effectively recommending action
Value to the users of knowing various evaluation outcomes	Far in excess of the costs of conducting monitoring and evaluation
Links to the direct intervention through which action based upon monitoring or evaluation information will come	Described, plausible and in existence

In an evaluable programme model, stakeholders are also likely to adopt recommendations for programme improvement and use the evaluation findings (Jung & Shubert, 1983). In addition, methodological requirements of the evaluation can be easily implemented (Rutman, 1980).

By working towards an evaluable programme model, the EA process therefore fosters a set of factors/conditions that determine the preparedness of a programme for useful evaluation. These factors and conditions can be conceptualised as evaluability parameters and determine the capacity to evaluate and the willingness to evaluate. In other words, they determine: (a) whether a programme has the necessary structural features or maturity for useful evaluation (e.g., an explicit and plausible programme theory, monitoring data); (b) whether it is feasible to implement the methodological requirements of the evaluation; (c) whether there is a need for a rigorous evaluation; and (d) whether programme stakeholders are willing to engage in the evaluation process and facilitate useful evaluation. Each of these factors are discussed below.

Well defined and plausible structural features.

The importance of having well-defined programme characteristics is succinctly captured in a statement made by Weiss (1972, p. 53):

“When programs are well-conceptualised and developed, with clearly defined goals and consistent methods of work, the lot of evaluation is relatively easy. But when programs are disorganised, beset with disruptions, ineffectively designed, or poorly managed, the evaluation falls heir to the problems of the setting”.

As emphasised earlier, a clearly defined programme is essential so that evaluation findings can be related to an identifiable intervention that was found to be effective or ineffective. Rutman (1980) argued that evaluations of undefined programmes are analogous to asking whether drugs work, without specifying the type of drug, the dosage, how it is administered, for which problem, and to what type of people.

A clear description of programme components also provides the basis for developing procedures to assess programme implementation. Assuming that a given programme has been implemented as per the original design can lead to erroneous conclusions about the effectiveness of the programme. Similarly, a set of clearly specified outcomes is a precondition for undertaking impact evaluations as it provides a basis for the development of appropriate (valid and reliable) measures.

For evaluations to be useful, the assumptions underlying the programme must also be plausible (i.e., there must be a realistic chance of attaining the specified outcomes). Evaluations that aim to determine the effectiveness of programmes whose outcomes are not realistic will predictably produce negative results (Rutman, 1980). Evaluators must be able to distinguish between rhetorical goals (often grandiose in nature) and plausible outcomes before developing procedures to measure their attainment. Two questions must be explored in determining the plausibility of outcomes:

- Does the programme actually direct efforts toward the stated outcomes?
- Does the programme make plausible causal assumptions about the problem it aims to solve?

Schmidt et al. (1979) proposed a framework that specifies the structural elements that must present and the requirements that must be satisfied for a programme description to be judged evaluable. If one or more of these elements are missing, and/or one or more of these requirements are not met, the likelihood that the evaluation will not be useful is higher. The different structural elements and accompanying requirements are presented in Table 2.

Table 2

Structural Elements and Requirements for an Evaluable Programme Description

Structural Elements	Requirements
Event sequence	Acceptable Well-defined
Event description	Acceptable Well defined Valid
Measure	Acceptable Well defined
Expected values	Acceptable Well defined Plausible
Evidence	Acceptable Well defined Cost effective Known to be reliable
Use of information	Acceptable Well defined Plausible

Each of the requirements articulated in Table 2 are described below:

- A description is acceptable when it is aligned to the expectations of policymakers.
- A description is valid when it accurately represents the programme activities, as implemented in practice.
- A description is considered plausible only when there is evidence to support its plausibility.
- The data system (defined in a description) is feasible when there are minimal cost or political constraints.
- The data system (defined in the description) is reliable when provisions are made for repeated observations and additional verifications prior to use.
- Stakeholders' expectations of the programme and the evaluation are plausible when these are in line with known resources and activities implemented.

Based on the above discussion, a well-defined programme (with clearly specified and plausible outcomes) increases the potential usefulness of an evaluation. However the final determination is based on the feasibility of measuring particular components or outcomes in a manner that will meet the evaluation's purposes (Rutman, 1980).

An overriding factor in determining the feasibility of conducting an evaluation, as discussed earlier, is the expected use of the information produced by the evaluation. It should be noted that: (a) the purpose of a given evaluation establishes the methodological requirements of the evaluation, and (b) evaluability relies, to a large extent on the feasibility of implementing those methodological requirements (Rutman, 1980). Evaluation purposes can be characterised as explicit or covert. Explicit purposes include searching for more cost-effective means of implementing a given programme, and demonstrating that the programme has been implemented as planned. In many instances programme evaluation simply serves a ritual to meet imposed requirements of funding agencies. In other situations however, evaluations are a means of making programmes look good, and delaying needed action to solve pressing problems (Rutman, 1980). These covert purposes can have a major influence on the feasibility of conducting a particular evaluation. For instance, programme personnel might be more resistant to implement a rigorous evaluation design when these covert purposes are dominant.

Table 3 summarises the potential purposes that an evaluation can serve.

Table 3

Potential Purposes of Evaluations

Purpose	Definition
Accountability	Compile data to demonstrate to stakeholders that the programme is functioning as expected
Monitoring	Routinely examine data to track expenditures, accomplishments, and other key indicators to guide programme management
Improvement	Identify operational strengths and weaknesses to devise ways to improve a programme
Understanding	Identify essential programme elements and opportunities for streamlining or enhancing the program
Replicability	Determine to what extent your programme can be well implemented in different settings
Judgement	Assess whether the programme provides a worthwhile return on investments of time, money, and other resources
Knowledge	Assess how effectively the programme is achieving its desired outcomes
Development	Exploring, building, and testing new ways to meet a target beneficiaries' needs

Note: Table adapted from Leviton et al. (2010).

Weiss (1998) succinctly summarised the evaluation purposes presented in

Table 3 along two dimensions: evaluation for decision making and evaluation as a tool for organizational learning. Decision making can revolve around the following aspects: (a) midcourse corrective actions (e.g., redefining the programme's eligibility criteria); (b) expanding the programme to new sites or discontinuing the programme altogether; (c) testing a new programmatic approach to service delivery and; (d) deciding whether to continue programme funding. Evaluation can serve as a tool for organizational learning by facilitating programme understanding, encouraging programme staff to reflect on the consequences of their work and ways in which they can improve their practice, and emphasising the link between desired programme goals and day-to-day activities.

Feasibility of implementing methodological requirements.

It can be argued that the greater the importance attached to evaluation findings, the higher the standards of precision and acceptable method/design required (Rutman, 1980). For example, more stringent methodological requirements are needed when evaluating large-scale programmes, which if proven effective, could influence national policy. Similarly, the purpose of an evaluation/stakeholders' demands determine to a large extent the methodological requirements of the evaluation. If the purpose of the evaluation is to test the programme for causal effectiveness, a number of criteria must be fulfilled to ensure that the methodological requirements of the evaluation can be implemented. Testing for causality involves a comparative assessment of the measured outcomes against an estimate of what those outcomes would have been in the absence of the programme. Such an assessment requires an experimental or quasi-experimental design (Shadish, Cook, & Campbell, 2002). Strong estimates of programme effects can be accomplished by randomising settings, outcome variables, and programme recipients/beneficiaries to different programme conditions (Reichardt, 2011).

Although randomised controlled trials (RCTs) are often considered as one of the strongest designs for unbiased estimates of programme effects, an extensive set of criteria has to be met for this design to be employed effectively (Patsopoulos, 2011; Victora, Habicht, & Bryce, 2004; White, Sabarwal, & de Hoop, 2014). These criteria include:

- A well-defined and clearly articulated programme model.
- Feasibility of collecting valid and reliable outcome measures.
- Evidence to support that the programme can realistically produce the desired outcomes.
- High implementation fidelity.
- Sufficient statistical power to detect the anticipated programme effect.
- Sufficient financial resources and evaluation expertise.

Additional considerations include: (a) a well-established participant recruitment and enrolment process; (b) an understanding of the characteristics of the target population, programme participants and programme environment; and (c) an adequate programme size. Each of these requirements is briefly discussed below.

An established participant recruitment and enrolment process.

The method of recruitment is an important consideration in designing an impact evaluation as it can point to potential sources of selection bias, dictate the feasibility of an experimental evaluation approach, and offer ways to derive a comparison group if a non-experimental approach is adopted. For these reasons, the recruiting methods must be thoroughly understood by the evaluator and must remain consistent throughout the evaluation process. The enrolment process is also important to consider because it may be a source of selection bias. If a given programme used a particular criterion to select participants such that those allowed to participate are most likely to experience successful outcomes, then not controlling for this selection will lead to an overestimation of the programme's effect.

An understanding of the characteristics of the target population, programme participants and programme environment.

Having an understanding of the characteristics of the target population, the characteristics of programme participants, and the economic and social environment in which the programme operates is important in designing an impact evaluation.

This information can assist the evaluator in developing the sampling methods to ensure that the sample is representative of the target population. An understanding of the characteristics of the population served and the programme context can also help evaluators interpret the findings once the evaluation has been conducted.

Adequate programme size.

In order to conduct an impact evaluation, there must be a sufficient number of individuals participating in the programme to obtain a reasonable level of statistical precision when estimating the programme's impact. The sample size necessary for conducting an evaluation will depend, in part, on the outcomes of interest and estimated programme impacts. For instance, the smaller the programme impact, the greater the sample size required to detect it.

Attempting impact evaluations where it is not appropriate, feasible, or affordable can lead to pseudo-impact evaluations (evaluations that do not provide information sufficiently robust to satisfy key stakeholders). What makes this undertaking complex and difficult are the constraints that compromise the implementation of the methodological requirements for such evaluations (Rutman, 1980). Political factors often limit the use of sampling to determine eligibility for the programme or to use random assignment procedures, and compromise the feasibility of implementing the rigorous evaluation designs due to pressing timelines. Similarly legal and ethical constraints limit the opportunity for implementing particular designs (those that for example require denial of service to a randomly chosen control group) and data collection procedures. The methodological requirements of a given evaluation design can also place additional demands on programme implementers (e.g., establishment of control groups, regular follow-up of programme beneficiaries/recipients to collect necessary data, and maintain programme records), thus justifying their resistance to the evaluation (Rutman, 1980).

Methodological requirements often intrude on the programme by making demands that affect programme delivery. For example, the evaluator might need to utilise programme staff for data collection purposes. It is therefore important to assess the feasibility of implementing the method/design that will provide information at the

required level of precision. According to Rutman (1980), the following questions must be addressed when determining the feasibility of implementing methodological requirements of a given evaluation:

- To what extent can information requirements be met at the desired degree of validity and reliability, considering cost and other types of constraints?
- To what extent will political factors undermine efforts to develop and implement valid and reliable measures?
- Are there major obstacles in obtaining the data? (e.g., ethical standards and administrative restrictions may limit access to data)
- What are the cost implications for obtaining the required information?

Figure 3, adapted from Davies (2013), summarises the different aspects discussed above and how they relate to evaluation design/methodological requirements.

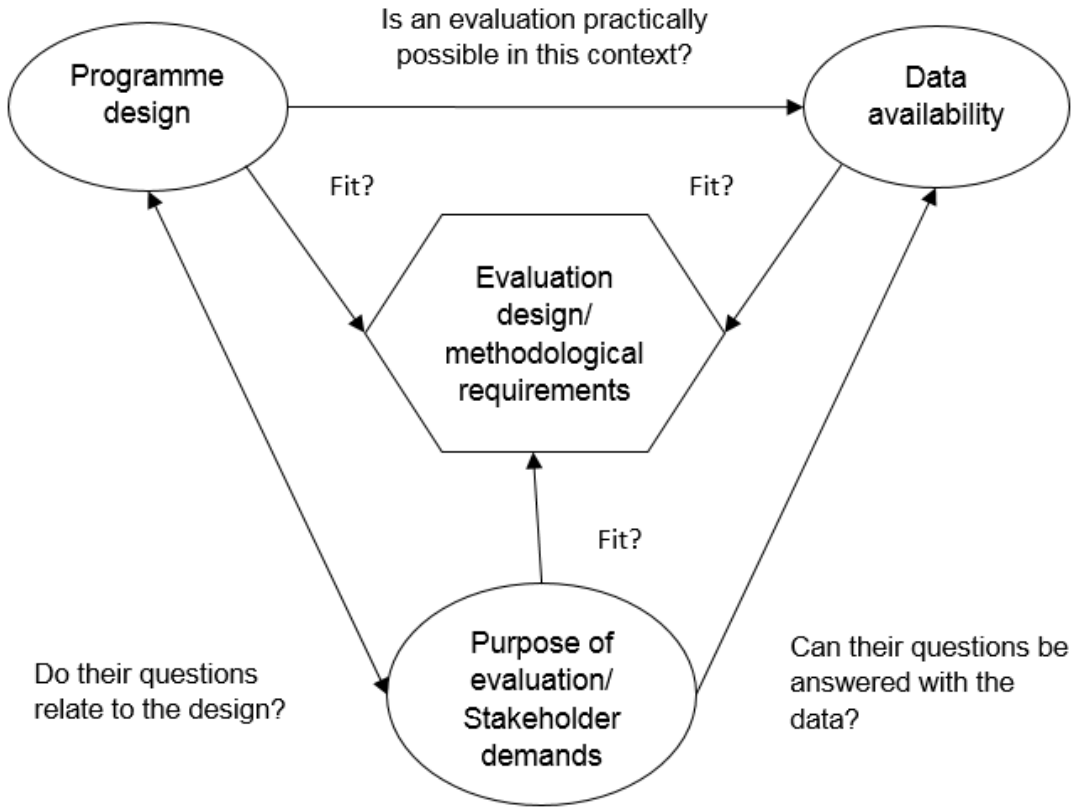


Figure 3. Assessing the feasibility of implementing methodological requirements

Need for a rigorous evaluation.

Smith (1981) argues that a programme might have the necessary structural attributes for a useful evaluation but this does not necessarily mean that an evaluation is warranted. When determining the need for an evaluation a number of factors have to be considered. These include: (a) the amount of resources that have been invested in programme design and implementation, (b) the potential social benefits of the programme and anticipated magnitude of these benefits, (c) the level of public interest in the programme, (d) the relevance of evaluative information to future policy formulation or programme decision, and (e) the need for accountability (Smith, 1981).

Stakeholders' willingness to engage in the evaluation.

A programme might warrant an evaluation but stakeholders might not be willing to engage in the evaluation process. This resistance can occur throughout the evaluation process, from the inception of an evaluation to the utilisation of its findings (Taut & Brauns, 2003). Stakeholders' resistance to evaluation can take many forms but usually involves a set of behaviours that aim to maintain the status quo, in the face of real or perceived pressure to change this status quo (Schwandt & Dahler-Larsen, 2006). These counterproductive behaviours in many cases result from evaluation anxiety, and include: (a) withdrawing from the evaluation process or refusing to work with evaluators, (b) attacking reasonable evaluation feedback and accusing evaluators of a hidden agenda, (c) ignoring well supported results, and (d) hiding programme weaknesses (Donaldson, Gooler, & Scriven, 2002). These counterproductive behaviours undermine the evaluability of a programme by limiting access to required information, compromising the quality of the data collected, and reducing the likelihood that evaluation findings will be utilised. By capitalising on stakeholder involvement, the EA process can shed light on disagreements among stakeholders and between stakeholders and evaluators. Unresolved disagreements might indicate that the programme is not ready for useful evaluation (Leviton et al., 2010). In addition, the EA process might reveal that programme stakeholders are reluctant to make critical changes to programme design and set stringent limits to the scope of the evaluation due to ideological or political reasons (Jung & Shubert, 1983;

Leviton et al., 2010). These conditions might indicate that the programme does not fit an evaluable programme model.

Other evaluability parameters.

Another issue that might signal that a programme is not ready for useful evaluation is when stakeholders and evaluators do not agree on the cost and timeline proposed in the EA report (even though the specified cost and timeline are reasonable and realistic given the desired scope and level of evaluation). The nature of the basic data requirements and data collection techniques affects evaluation costs to a large extent. For instance, personal interviews and observations are more expensive than self-report questionnaires and use of secondary data. The degree of precision required (e.g., need for multiple measures/triangulation of data and a large sample size) also has a major bearing on the costs of an evaluation (Rutman, 1980). If stakeholders are reluctant or unable to allocate required resources to the evaluation, the evaluability of the programme is questionable (Leviton et al., 2010). Budget and time constraints might compromise the quality of the evaluation (for example information needs are not adequately and validly addressed) and hence limit the usefulness of the evaluation (Bamberger, Ruth, & Mabry, 2006). The evaluability of a programme might be further compromised if the available evaluation capacity and expertise is not in line with the scope and level of evaluation specified in the EA report.

Assessment of Evaluability and Decision to Evaluate or not

It is important to distinguish between an evaluator's assessment of evaluability and his/her decision to accept or decline an evaluation contract. These two judgements might not necessarily follow from one another, or relate to one another. For example, evaluability might be a necessary but not sufficient condition for accepting an evaluation contract. Evaluation is a market-based profession and economic considerations inevitably play an important role in evaluators' decision to accept or decline an evaluation contract (Smith, 1998).

Smith (1998) argues that the evaluation profession serves two purposes, namely guild maintenance and societal improvement. Since the profession serves, in part, to protect and promote the livelihood of its members, one can argue that if an evaluation is not profitable to the evaluator, the evaluation contract should not be accepted. A profitable contract does not necessarily refer to one with high financial gains, but one that might promote the evaluator's career and reputation, or provide the evaluator with the opportunity to work with valued clients or in a desirable geographical setting.

An evaluator can easily gauge the monetary benefit of an evaluation contract and use this as a basis for deciding whether to accept or decline the contract. Assessment of evaluability, on the other hand, is more complex and involves an aggregation of judgements across multiple evaluability dimensions. An evaluation contract might not always simultaneously satisfy conditions of economic self-interest and evaluability. A decision was therefore taken to separate these two issues and keep indicators of economic self-interest neutral in order not to contaminate evaluators' assessment of evaluability. Of particular interest in this study, is evaluators' assessment of evaluability independent of their decision to accept or decline an evaluation contract.

A Framework for Evaluability based on the Literature

The literature on evaluability, more specifically the implicit and explicit evaluability criteria identified, has been summarised in Appendix A. The following broad evaluability dimensions emerged consistently in the literature:

1. Programme objectives (Chen, 2005; Rutman, 1980; Schmidt et al., 1979; Stenberg, 1983; Weiss, 1998; Wholey, 1979, 2010).
2. Programme data (Chen, 1995; Kaufman-Levy & Poulin, 2003; Schmidt et al., 1979; UNIFEM, 2009; Wholey, 1979, 2010).
3. Programme theory (Chen, 1995; Horst et al., 1974; Jung & Shubert, 1983; Leviton, 2010; Rog, 1997; Schmidt et al., 1980; Stenberg, 1983; Taut & Brauns, 2003).

4. Programme design (Davies, 2013; Kaufman-Levy & Poulin, 2003; Horst et al., 1974; UNIFEM, 2009; Wholey, 1979, 2010).
5. Programme implementation (Kaufman-Levy & Poulin, 2003; Wholey, 1979, 2010).
6. Stakeholder willingness (Kaufman-Levy & Poulin, 2003; Horst et al., 1974; Newcomer et al., 2010; Rutman, 1980; UNIFEM, 2009; Wholey, 1979, 2010).
7. Purpose of evaluation (Newcomer et al., 2010; Rutman, 1980; Weiss 1998).
8. Evaluation budget, timeframe, and methodological requirements (Davies, 2013; Newcomer et al., 2010; Rog, 1997; Smith, 1981; UNIFEM, 2009; Wholey, 1979, 2010).
9. Evaluation capacity and expertise (Rog, 1997; Smith, 1981; Newcomer et al., 2010).

I developed a programme evaluability framework (see Figure 4) by collapsing the above dimensions into four distinct categories: (a) programme characteristics/programme structural features (dimensions 1-5), (b) stakeholder characteristics (dimensions 6 and 7), (c) logistical requirements (dimension 8), and (d) evaluator characteristics (dimension 9).

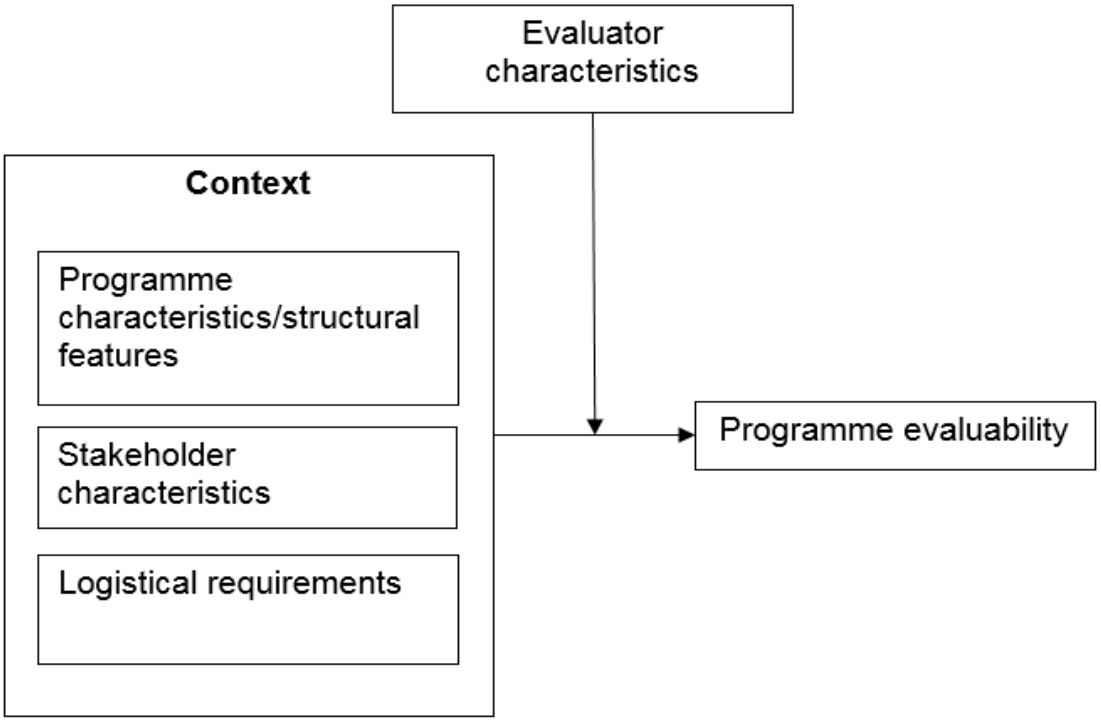


Figure 4. Programme evaluability framework.

Programme characteristics, stakeholder characteristics and logistical requirements are embedded in the programme context/reality and can be conceptualised as independent variables that influence evaluators' assessment of programme evaluability. Evaluator characteristics, on the other hand, are independent of the context, and can be conceptualised a moderating variable.

The evaluability framework developed for the purpose of this study consolidates Davies's (2013) work on evaluability. As part of a working paper commissioned by the UK Department for International Development (DFID), Davies (2013) examined guidance documents produced by the eight international development agencies, and developed a checklist measuring three core evaluability dimensions (programme design, availability of information, and institutional context). The evaluability framework presented in Figure 4 was derived from a comprehensive review of the literature on evaluability, and integrates the work of multiple prominent evaluation theorists and international development agencies.

Operationalising Programme Evaluability Criteria

It is clear that the concept of evaluability has not been clearly articulated in the literature. An attempt has however been made to systematise and integrate the different evaluability parameters that have been implicitly or explicitly articulated in the literature. The framework presented in Figure 4 warrants empirical validation as it is unclear whether it captures all the relevant evaluability criteria that practitioners actually use in practice. The extent to which practitioners use the specified criteria is also unclear. For such an investigation to be possible the evaluability criteria presented in Figure 4 must first be standardised (conceptually) and operationalised. Table 4 presents the disaggregated evaluability criteria specified by various authors. These have been derived from Appendix A.

Table 4

Disaggregated Evaluability Criteria and Requirements

Disaggregated Evaluability Criteria	Requirements	Authors
Programme characteristics/structural features		
Programme objectives/goals/outcomes	Well-defined/clearly specified	Chen (2005); Horst et al., (1979); Rutman (1980); Stenberg (1983)
	Realistic/Plausible	Chen (2005); Rutman (1980); Schmidt et al. (1980); Stenberg (1983)
	Measurable	Kaufman-Levy and Poulin (2003); Stenberg (1983); Wholey (1979, 2010)
	Agreed upon	Stenberg (1983); Wholey (1979, 2010); Weiss (1998)
Programme data	Adequate	Chen (2005)
	Easily obtainable/accessible	Chen (2005); Kaufman-Levy and Poulin (2003); Schmidt et al. (1980); UNIFEM (2009); Wholey (1979, 2010)
	Reliable	Schmidt et al. (1980)
Programme theory	Documented	Horst et al. (1979); Kaufman-Levy and Poulin (2003); Wholey (1979, 2010); UNIFEM (2009)
	Plausible	Kaufman-Levy and Poulin (2003); Rutman (1980); Schmidt et al. (1980); UNIFEM (2009); Wholey (1979, 2010)

Table 4 (cont.)

Disaggregated Evaluability Criteria and Requirements

Disaggregated Evaluability Criteria	Requirements	Authors
Programme characteristics/structural features		
Programme design	Clearly defined intervention	Horst et al. (1979); Schmidt et al. (1980)l UNIFEM (2009);
	Clearly defined target beneficiaries	Kaufman-Levy and Poulin (2003);UNIFEM (2009)
Programme implementation	Implemented as intended	Kaufman-Levy and Poulin (2003); Rutman (1980); Wholey (1979, 2010)
Stakeholder characteristics		
Willingness	Willingness to facilitate evaluation process	Stenberg (1983)
Authority	Authority to facilitate evaluation process and act on evaluation findings	Horst et al., (1979); Jung and Shubert (1983); Leviton (2010); Taut and Brauns (2003);
Transparency	Clearly identified information needs	Chen (2005); Rutman (1980);Wholey (1979; 2010);
Logistical requirements	Feasibility of implementing desired methodology	Rutman (1980)
	Level of evaluation feasible	Strosberg & Wholey (1983)
	Adequate budget and timeline	Bamberger et al.(2006); Newcomer at al. (2010);

Different evaluators might articulate the minimum evaluability criteria and interpret the associated requirements differently. For example, some evaluators confuse programme goals with objectives and outcomes, when the distinction between these concepts is critical (Rossi, Lipsey, & Freeman, 2004). In an attempt to standardise the interpretation of the minimum evaluability criteria and requirements, key definitions from Rossi et al. (2004) are presented below (where necessary). Since its first publication in 1979, Rossi et al.'s evaluation textbook, *Evaluation: A systematic approach,* has been a benchmark text in the field of evaluation. The seventh and latest edition of the textbook was published in 2004 and was used to derive the key definitions relating to programme characteristics/structural features. Some of the definitions presented have been derived from the glossary and are quoted in full. Other definitions have been derived from the relevant chapters.

After providing the standardised definition from Rossi et al., operational definitions of the evaluability criteria will be provided. According to Crano and Brewer (2002, pp.9) operationalisation, or "the translation of conceptual variables into scientifically researchable variables" involves two steps: (1) a redefinition of an abstract concept into something that is "observable or manipulable", and (2) "a specification of the procedures and instruments required to make the actual observation." The first step will be discussed below and the second step will be discussed in full in the method section.

Evaluability Criterion 1: Programme Characteristics/Structural Features

Programme objectives/goals/outcomes.

A programme goal refers to a "statement, usually general and abstract, of a desired state toward which a programme is directed" (Rossi et al., 2004, p. 431). A programme objective, on the other hand, is a "specific statement detailing the desired accomplishments of a programme together with one or more measurable criteria of success" (Rossi et al., 2004, p. 432). A programme outcome refers to "the state of the target population or the social conditions that a programme is expected to have changed" (Rossi et al., 2004, p. 429).

From the standardised definitions provided above, it is clear that programme goals need to be disaggregated further from programme objectives/outcomes as they refer to different concepts. However, based on the definitions again, it would seem as if objective and outcomes refer to the same concept, namely intended or actual change in the target population. Based on pragmatism and common usage, a decision has been taken to use the label*, outcome*, and discard the label, *objective*. This means that we are dealing with two separate concepts here, namely programme goals and programme outcomes.

For programme goals and outcomes to be clearly specified, they need to be "stated in sufficiently clear and concrete terms to permit a determination of whether they have been attained" (Rossi et al., 2004, p.157). For programme goals and outcomes to be realistic or plausible, they have to be within "influence of the programme" (i.e., the programme can be reasonably expected to produce desired goal or outcome). For programme goals and outcomes to be measurable, they have to be associated, at least implicitly, with some criteria by which performance can be judged. Stakeholder consensus on what the programme is trying to achieve would indicate agreed-upon programme goals and outcomes.

As indicated above, programme goals and outcomes are often used in conjunction with specific requirements for evaluability. Based on this evidence, the following four evaluability criteria can be operationalised within this category:

- Programme goals are clearly specified.
- Stakeholders agree on programme goals.
- Programme outcomes are realistic.
- Programme outcomes are measurable.

Programme data.

Programme data usually reflect information about the programme's implementation process and its progress in terms of its outcomes. A well-maintained and updated record system might be an indication of adequate and reliable programme data. That

is, the programme has procedures and systems in place to track implementation of programme activities (process) and progress toward intended results (outcome). Apart from the existence of such procedures and systems, it is also easy for evaluators to access the data on the data systems.

Based on the above definitions, it is clear that programme data are often used in conjunction with specific requirements for evaluability, such as adequacy, reliability and accessibility. The following three evaluability criteria can therefore be operationalised within this category:

- Programme data are adequate.
- Programme data are reliable.
- Programme data are easily accessible.

Programme theory.

Programme theory refers to “the set of assumptions about the manner in which a programme is expected to produce desired outcomes and the strategy and tactics the programme has adopted to achieve its goals and objectives” (Rossi et al., 2004, p. 432). A programme theory can be implicit or explicitly documented. Rossi et al. (2004, p. 135) indicated that “the first step in assessing a programme theory is to articulate it”. Once articulated, evaluators can make a judgment about the programme theory’s plausibility (i.e., whether or not the cause and effect links in the document are plausible)

As indicated above, the two evaluability requirements most commonly used in conjunction with programme theory, are explicitly stated (documented) and plausible. The following two evaluability criteria can therefore be operationalised within this category:

- Programme theory is explicitly stated.
- Programme theory is plausible.

Programme design.

Programme design encompasses programme operations (the service delivery system) and the population it serves. Programme operations are clearly defined when they are specified in concrete terms (e.g., “the programme has such and such resources, facilities, personnel, and so on, [and is] organised and administered in such and such a manner, and engages in such and such activities and functions…”) (Rossi et al., 2004, p. 141). Target beneficiaries are clearly defined when the definition of the target population permit targets to be distinguished from non-target units in a relatively unambiguous and efficient manner” (Rossi et al., 2004, p.118).
From the above definitions, it is clear that: (a) programme design incorporates both service delivery operations and target beneficiaries; and (b) the same evaluability requirement, clear definition, is often used in conjunction with each component. The following evaluability criteria can therefore be operationalised within this category:

- Service delivery is clearly defined.
- Target beneficiaries are clearly defined.

Programme implementation.

Programme implementation encompasses service utilisation and programme organization. A programme is considered to be implemented as planned when “the programme [adequately] performs the activities specified in the programme design that are assumed to be necessary for bringing about the intended social improvements” (Rossi et al., 2004, p. 171).

While the above definitions point to the complexity of programme implementation, a decision has been taken not to disaggregate the construct into the two separate components of service utilisation and programme organization, as programme organization is included under programme design. As such, the evaluability criterion *programme is implemented as intended* can be operationalised within this category:

.

Evaluability Criterion 2: Stakeholder Characteristics

Programme stakeholders are “individuals, groups, or organizations having a significant interest in how well a programme functions, for instance, those with decision-making authority over the programme, funders and sponsors, administrators and personnel, and clients or intended beneficiaries” (Rossi et al., 2004, p. 435). Not all programme stakeholders are equally important or involved in a given evaluation scenario. The evaluation sponsor typically has the most influence in the process in terms of initiating and commissioning the evaluation, specifying how and when it will be conducted, and deciding who will conduct the evaluation (Rossi et al., 2004). However, programme administrators and personnel may also influence the extent to which the evaluator can gain access to required data. A decision was therefore taken to retain the label, *stakeholder*, without defining it too strictly.

Three dimensions of stakeholder characteristics are specified in Table 4, namely: willingness, authority, and transparency about the purpose of the evaluation. Stakeholder willingness refers to the way in which the stakeholder collaborates with the evaluator (Rossi et al., 2004, p.49). Stakeholder authority relates to the authority to facilitate the evaluation process and act on evaluation findings (Jung & Shubert; 1983; Horst et al., 1974; Leviton; 2010; Taut & Brauns, 2003). No specific evaluability requirement is used in conjunction with these two criteria in the literature, but one could envisage the level of motivation to exercise collaboration or authority.

Stakeholder transparency relates to the purpose of an evaluation, more specifically the extent to which this purpose is defined and communicated. Rossi et al. (2004) does not specifically discuss evaluation purpose or stakeholder transparency. However, Weiss (1998) has written extensively on this issue. She warns evaluators to be alert to the fact that sometimes the purpose of an evaluation may be unacknowledged and covert (e.g. a non-legitimate reason for requiring an evaluation). However, often evaluation purposes are overt and acknowledged and such purposes are usually decision-making or organizational learning. It can be assumed that legitimate purposes are more likely to facilitate a smooth evaluation than political purposes. Legitimate purposes may contribute to improved programme understanding or rational decisions regarding programme revision or closure.

Weiss (1998) argues that evaluation is inherently tied to the principle of utility. If an evaluation does not contribute to decision making or organizational learning, it is best described as an exercise of futility. For example, when evaluation is used as a delaying tactic or a public relation tool, it loses its intended legitimacy. Understanding the motives of the evaluation sponsor and clarifying the expectations of this stakeholder group is critical. This entails the identification and prioritisation of information needs.

Based on the above, three evaluability criteria can be operationalised within the category of stakeholder characteristics:

- Stakeholders are willing to collaborate with the evaluator.
- Stakeholders have authority to act on evaluation findings.
- Stakeholders are transparent about the purpose of the evaluation.

Evaluability Criteria 3: Logistical Requirements

Three dimensions of logical requirements are specified in Table 4, namely: feasibility of implementing the desired methodology, feasibility of conducting the desired level of evaluation, and adequate budget and timeline. Rossi et al. (2004) refer to programme evaluability but not to these four specific criteria. For this reason, the work of the authors in Table 4 will be used to standardise these criteria.

As indicated earlier in this chapter, Rutman (1980) indicated that many evaluations are of limited value due to methodological flaws. It could be assumed that a request for an evaluation should take into account the time, effort and cost of rigorous methods.

Apart from methodological rigour, the scope and level of the evaluation must be feasible within the specified timeframe (Strosberg & Wholey, 1983). A complex outcome or impact evaluation will require a longer timeframe than a simple process evaluation.

Adequate budget and timeline are discussed by a number of authors. In fact, budget and time constraints underlie Bamberger, Rugh, Church, and Fort's (2004) shoestring approach to evaluation. This approach was designed to assist evaluators operating under those constraints to conduct evaluations with the maximum possible methodological rigor by, for example, simplifying the evaluation design, revising the sample size, and exploring economical data collection methods.

Based on Table 4 and the discussion above, the following evaluability criteria can therefore be operationalised with the category logistical requirements:

- Budget is adequate for the evaluation.
- Timeframe is adequate to complete the evaluation.
- Type of evaluation required (process, outcome or impact) is feasible.
- Required evaluation methodology is feasible.

It is clear that the concept of evaluability is vague and ambiguous. Tighter definitions of the concept and concrete criteria that can be used to assess evaluability have been presented in this chapter. An integrated model for evaluability was derived based on the literature reviewed. Each of the four blocks of the model was disaggregated into 18 distinct evaluability criteria and presented in Table 4. Standardised conceptual definitions for each of these criteria were then provided using key definitions from Rossi et al. (2004). The list was further refined based on pragmatism, common usage and the conceptual definitions provided. All the operationalised evaluability criteria are presented in Table 5.

Table 5

Operationalised Evaluability Criteria

Category	Criteria
Programme characteristics	
Programme goals and outcomes	Programme goals are clearly specified
	Programme outcomes are realistic
	Programme outcomes are measurable
	Stakeholders agree on programme goals
Programme data	Programme data are adequate
	Programme data are reliable
	Programme data are easily accessible
Programme theory	Programme theory is explicitly stated
	Programme theory is plausible
Programme design	Service delivery is clearly defined
	Target beneficiaries are clearly defined
Programme implementation	Programme is implemented as intended
Stakeholder Characteristics	Stakeholders are willing to collaborate with the evaluator
	Stakeholders have authority to act on evaluation findings
	Stakeholders are transparent about the purpose of the evaluation
Logistical Requirements	Budget is adequate for the evaluation
	Timeframe is adequate to complete the evaluation
	Type of evaluation required (process, outcome or impact) is feasible
	Required evaluation methodology is feasible

Operationalising Evaluator Characteristics

In the model presented in
Figure *4*, evaluator characteristics consist of level of experience, qualifications/training, and practice context. These variables were operationalised using relevant participant demographics items embedded in an online questionnaire. The questionnaire will be described in full in Chapter 4.
The conceptual and empirical link between these evaluator characteristics and evaluation practice is discussed below.

Experience.

Evaluators draw on conceptual resources and practical knowledge to approach and solve problems, and make decisions in a real evaluation context (Donaldson & Lipsey, 2008; Schwandt, 2007; Tourmen, 2009). Tourmen (2009) argues that the type and level of conceptual resources an evaluator has are, in turn, are related to his or her level of experience. Furthermore, experienced practitioners accumulate pragmatic knowledge "that is different, but not completely disconnected from formal theoretical knowledge in evaluation" (Tourmen, 2009, pp. 8).

Some studies have shown that experienced practitioners are more able to anticipate and take into account situations when making decisions, compared to novice practitioners. Using a unique method, framed around theories of activity and ergonomics, Tourmen (2009), for example, found that compared to novice practitioners, experienced evaluators: (a) make broader and more nuanced diagnosis of the evaluation context; (b) question evaluation requirements, clarify and negotiate stakeholders' expectations, and transform the evaluation situation in a more active way; (c) make more predictions about the evolution of the evaluation context; and (d) make more compromises between conflicting evaluation goals. Tourmen (2009) attributed these findings to differences in cognitive resources and conceptualisation of situations and argues that evaluators learn to integrate formal knowledge into their own conceptualisations with experience–a process that Vergnaud, Pastré, and Mayen (2006) refer to as the *pragmatisation process*.

In order to operationalise experience, the number of evaluations completed in the last five years and participants' self-assessment of their level of experience in conducting different types of evaluation were utilised.

Qualification/training.

A number of taxonomies of evaluation skills have been proposed by evaluation theorists and practitioners (e.g., King, Stevahn, Ghere, & Minnema, 2001; McGuire & Zorzi, 2005) to describe the necessary competencies required to conduct technically challenging evaluations. These authors have emphasised that is it critical for evaluators to know about the different evaluation concepts, methods, and tools in order to overcome the methodological prejudice of their primary discipline and select the most appropriate tools in relation to a particular evaluation context.

This knowledge can be acquired through different forms of qualifications or training. It should be noted that programme evaluation is not a statutory profession and therefore does not have minimum requirements of qualifications for its members. Entry into the profession is not controlled (Engle, Altschuld, & Kim, 2006). In fact programme evaluators come from various disciplines (Lavelle & Donaldson, 2015) and build their technical expertise through training offered by evaluation capacity-building organizations, such as InsideOut, Southern Hemisphere, and Impact Consulting in South Africa, and professional associations, such as the American Evaluation Association (AEA) in the United States of America. University-based evaluation training is also offered in the form of fully-fledged evaluation programmes or isolated modules. This will be discussed in more depth in Chapter 3.

One can reasonably argue that university-based evaluation training offered at postgraduate level in evaluation specific programmes, has the potential to enhance the technical competence of individuals designing and conducting evaluations (Lavelle & Donaldson, 2010). This effect might be more pronounced for evaluation professionals who conduct technically demanding evaluations as their primary job responsibility. Qualification will be operationalised in terms of the level of tertiary education (i.e., Doctoral, Master's, or Bachelor's degree). A further area of

operational interest was the discipline in which the degree was obtained. This could be either in programme evaluation or not.

Practice context.

In Chapter 3, the selection of four different countries where the practice of evaluation could be located is discussed in detail. When the choice of country is operationalised, the following categories are used:

- Developed or developing country.
- Strong or weak evaluation cultures and capacity.
- Early adopter or late adopter of evaluation practice.

The operational definition of each evaluator characteristic discussed above is presented in Table 6.

Table 6

Operational Definition of Evaluator Characteristics

Evaluator Characteristic	Definition
Experience	Number of evaluations completed in the last five years
	Self-assessment of level of experience in conducting different types of evaluations
Qualification	Level of tertiary education
	Discipline in which degree was obtained
Practice context	Developed or developing country, or both
	Country with strong/weak evaluation culture and capacity
	Early adopter or late adopter of evaluation practice

Conclusion

Evaluability is not an absolute condition. Rather, it occurs along a continuum from more to less evaluable. The mix of evaluability criteria presented in this chapter can be used to assess whether a given programme is more or less evaluable. Despite the lack of systematic and concrete indicators in the literature, experienced evaluators however seem to know almost intuitively how evaluable a specific programme is. This experience guides their decisions in taking on evaluation contracts and eventually helping them produce high-level impact evaluations. At this stage we do not know what heuristics they use to assess evaluability. The purpose of this study is to investigate which criteria experienced and novice evaluators prioritise when assessing the evaluability of a programme and deciding whether to conduct an evaluation. Other than experience, the context within which evaluation decisions are taken might influence evaluability. For instance, in the United States of America (USA) and the United Kingdom (UK), programme evaluation is a mature and established science practised in developed economies. On the other hand, in developing economies like South Africa (SA) and Brazil, programme evaluation as a science and practice might look different. In this study, practice context will be used as one of the evaluator characteristics which might predict evaluability decisions. This variable is discussed in depth in Chapter 3.

CHAPTER THREE

COUNTRY OVERVIEW

Theorists have long recognised the role that context plays in shaping evaluation practice (e.g., Stake, 1990; Stufflebeam, 1971). In Chapter 2, the possible link between decision-making context (i.e., practice context) and how evaluators approach the assessment of evaluability was highlighted. Despite the growing emphasis on the role of context in evaluation, there is no unified understanding of what context means and how exactly it influences evaluation practice (Dahler-Larsen & Schwandt, 2012; Fitzpatrick, 2012; Rog, 2012). Greene (2005), for example, distinguishes between five dimensions of context, namely: (a) demographic characteristics of the setting and people in it, (b) economic features, (c) institutional and organizational climate, (d) interpersonal dimensions or typical means of interaction and norms for relationships in the setting, and (e) political dynamics of the setting. Authors such as Chouinard and Cousins (2009), on the other hand, focus on the cultural dimension of context. Of particular interest in this study is the practice of evaluation in different countries. Nevo (1982) first called evaluators' attention to contextual differences across countries and the questionable applicability of US-derived evaluation theories and models in other countries. Patton (2007, as cited in King & Greenseid, p. 12) further argued that "international diversity in evaluation practice is challenging our thinking about what constitutes good evaluation work".

The context of a country, the nature of its government, and expectations of NGOs and multinational development agencies are influences beyond a specific programme that might affect an evaluator's approach to evaluation (Fitzpatrick, 2012). The context of developing countries is different from that of developed countries in a number of ways. One can therefore reasonably assume that the practice of evaluation is different in developing and developed countries. Evaluators practising in developing countries and those practising in developed countries might prioritise a different set of evaluability criteria. Differences in practice might be more pronounced for evaluators who practise in countries with weak and emergent evaluation cultures and those who practise in countries with strong and mature evaluation cultures.

As highlighted in Chapter 2, the practice of evaluators from four different countries were examined. These four countries were selected on basis of their level of development, the strength of their evaluation culture and capacity, and whether they were an early adopter or late adopter of evaluation practice. In addition, the relative ease of accessing the target population through well-established evaluation associations in these four countries was also considered. The countries of interest to this study and their respective profile are presented in Table 7.

Table 7

Country Profile

Country	Level of development	Strength of evaluation culture and capacity	Evaluation Practice
Brazil	Developing	Weak	Late adopter
South Africa	Developing	Weak	Late adopter
USA	Developed	Strong	Early adopter
UK	Developed	Strong	Early adopter

As noted in Nielsen (2011), there is no standardised development taxonomy or development threshold that can be used to categorise countries based on their level of development. International development organizations such as the United Nations Development Programme (UNDP), the World Bank, and the International Monetary Fund (IMF) have different country classification systems. The UNDP's country classification system is, for example, built around the Human Development Index (HDI) and distinguishes between low, medium, high, and very high human development countries based on three indices, namely: longevity, income, and education. The World Bank's country classification system, on the other hand, distinguishes between low income, middle income, and high income countries based on Gross National Income (GNI) per capita. Within the middle income category countries are either classified as lower or upper middle-income. Low income and middle income countries are sometimes referred to as developing economies in the World Bank classification system. The use of the term is convenient and does not

imply that all countries in the low income and middle income categories experience similar development challenges (The World Bank, 2015).

The World Bank country classification system was used to categorise the four countries presented in Table 7 in terms of their level of development. As per the classification thresholds, SA and Brazil are upper middle income countries (developing countries) and the USA and the UK are high income countries (developed countries). What follows is an in-depth analysis of the four countries of interest, with a particular emphasis on the evolution and maturity of the field of programme evaluation in each country.

The development of the evaluation function in each country will be examined in relation to its past and current socio-political context as changes in a country's socio-political context dictate the evolution and maturity of the field (Neirotti, 2012). As such, this chapter first outlines the key historical and political markers, and the current socio-economic standing and challenges of the four countries of interest to this study. This is followed by an overview of the history and development of evaluation in these four countries. This section will cover four broad aspects: (1) the emergence and development of programme evaluation in each country, (2) key players/departments involved in advancing the monitoring and evaluation (M&E) agenda in each country, (3) evaluation capacity-building and training options in each country, and (4) the current challenges encountered by each country while advancing the M&E agenda. The organization of the sub-sections on the USA and the UK deviates slightly from that of Brazil and SA. An assessment of maturity of the discipline in these two developed countries and their major contributions towards the discipline will also be presented. This chapter will culminate into a discussion of the rationale for using Brazil, SA, the USA, and the UK as units of analysis in this study.

BRAZIL

Brazil's Key Historical and Political Markers

Brazil has experienced dramatic changes in its political landscape. In less than 200 years of independent history, Brazil has been a monarchy (1822–1889), an oligarchic republic (1889–1930), an authoritarian civil state (1930–1945), an autocratic democracy (1945–1962), a parliamentary democracy (1962–1964), an authoritarian military state (1964–1985) and finally a liberal democracy, fully established in 1988 (Mohanty, Thompson, & Coelho, 2011). Such drastic institutional changes affected the proper implementation of civil, social and political rights.

Two important transitions in Brazil's recent history is the coup d'état in 1964, following a brief democratic window, and the consolidation of the liberal democratic government over the period 1989-2002 (Codato, 2006). The decades following the establishment of the military dictatorship were marked by fierce suppression of political opposition, a centralised economy, and a significant increase in income inequality (Mohanty et al., 2011). However, a distinguishing feature of the Brazilian dictatorship was that it maintained some political institutions from the previous democratic regime. The Federal Congress continued to function and subsidiary elections for states and municipal governments, though limited, were allowed throughout the military years. In addition, social benefits were expanded to previously excluded sectors of the population. Housing, basic sanitation and several social assistance programmes were implemented in 1970s to increase state control over rural areas (Arretche, 2002).

This picture gradually started to shift in the last 27 years, with the end of the military rule and the establishment of the democratic constitution in 1988. The new constitution was guided by the principles of institutional decentralisation and popular participation. General elections were re-introduced and major macro-economic reform, which helped Brazil regain stability and control over inflation, were also initiated. Furthermore, the social policy arena was largely transformed with the introduction and expansion of policies for poverty reduction, increased public

participation in decision-making and greater integration of previously marginalised groups (Arretche, 2002).

Over the past few decades Brazil has made major strides in its efforts to address a number of socio-economic challenges and recover from its military dictatorship that lasted 21 years. Brazil is now South America's most influential country and one of the world's biggest democracies. It is one of the rising economic powers that forms part of the BRICS nations. BRICS is a grouping acronym that refers to Brazil, Russia, India, China and SA. These five countries are deemed to be at a similar stage of newly advanced economic development.

Brazil's Current Socio-Economic Standing

Brazil's gross domestic product (GDP) is US $1.775 trillion, which situates it in the upper middle income country category according to the World Bank classification (The World Bank, 2016a). The country is currently experiencing a deep recession - the GDP has contracted by 3.8% in 2015. The country's rapid economic and social progress over the period 2003- 2014 has however lifted 29 million people out of poverty and closed the inequality gap significantly (the Gini coefficient fell by 11% in the same period, as per The World Bank's current statistics). While Brazil experienced lower growth rates in the last decade than under the military regime, basic social indicators improved faster than in most other countries with similar income levels (Bertelsmann, 2016). For instance, extreme poverty (US$1.25 per day) dropped dramatically, from 10% in 2004 to 3.6% in 2012. Brazil is now close to achieving universal basic education and improved health in the whole country. In addition to a significant expansion in basic school enrolment (from 80% in 1980 to over 97% in 2002), school attendance rose from 85% to 97% in the period 1992-2008, followed by an increase in grade promotion rates (Bruns, Evans, & Luque, 2012). The health of Brazilians has also improved significantly on a number of levels, following drastic heath system reforms. Child mortality declined from 26 to 16 per 1,000 live births in the whole country from 2003 to 2015 (The World Bank, 2016a). Mortality from infectious diseases has declined substantially and the number of new HIV/AIDS cases has stabilised. The situation of young children in Brazil has also improved significantly in terms of child outcomes (Evans & Kosec, 2012). These

gains can be observed across sectors, including education, health, legal protection, water, and sanitation.

Brazil's overall socio-economic progress can be attributed to a number of factors, including its innovative approach to welfare, minimum wage policies, pension schemes linked to inflation, and well-focused social programmes (Lopez-Calva & Rocha, 2012; The World Bank, 2016b). Public investment in programmes for children and adolescents in Brazil, for instance, doubled between 2006 and 2009 (Evans & Kosec, 2012). The Bolsa Família conditional cash transfer program, for example, is one of the most effective and high profile social protection interventions in the world (The World Bank, 2016c). Other programmes include the Second Family Health Extension Programme, which provides medical attention to vulnerable groups that do not have easy access to the hospital system. Vaccination coverage and child nutrition have improved significantly following the implementation of this programme. Brazil's strategic partnership with the World Bank has also contributed to the improvement in social indicators and to the expanded access of the poor to basic infrastructure (The International Bank for Reconstruction and Development, 2004).

Brazil's Current Socio-Economic Challenges

As outlined in the previous section, there have been sustained improvements in Brazil's education system, both in terms of access and student learning. The education system in Brazil is no longer considered as one of the worst performing systems in upper middle-income countries. Despite having reached universal coverage in primary education, Brazil is now struggling to improve the quality and outcome of the system, particularly at primary and secondary levels (Bruns et al., 2012; The World Bank, 2016b). The education system is still not on par with other upper middle-income countries in terms of average learning levels, secondary education completion rates, and teacher quality, particularly in disadvantaged communities. While access to education has become more equitable over the past 15 years, there is still a significant gap between rich and poor in learning levels and completion rates.

With regard to early childhood care and education, Brazil has made significant progress in expanding the number of early childhood development (ECD) establishments and enrolments. However, the infrastructure of ECD establishments and the quality of care they offer are still deficient in a number of states, with some requiring extensive expansions to achieve the intended universal pre-school coverage and quality by 2016 (Evans & Kosec, 2012; Neri & Buchmann, 2007). In addition, there are still stark disparities in terms of access. While ECD education is practically universal for children from high income families, access to nurseries, in particular is considerably restricted for children from disadvantaged backgrounds.

As outlined in the previous section, income inequality in Brazil has fallen steadily over the last fifteen years, reaching a Gini coefficient of 0.5 in 2015 (The World Bank, 2016b). Despite these achievements, inequality remains relatively high for an upper middle income country and according to regional and international standards. In fact, Brazil's level of income inequality remains among the world's highest, and have shown signs of stagnation since 2015 (The World Bank, 2016b). In addition, while the rate of poverty declined significantly, there are many rural areas in Brazil that are afflicted by extreme poverty. As result, Brazil experiences extreme regional differences, in terms of access to basic services, and performance with regard to social indicators such as health, infant mortality and nutrition. The richer south and southeast regions perform better in terms of socio-economic indicators than the poorer north and northeast regions (The World Bank, 2016b; Lopez-Calva & Rocha, 2012).

Brazil’s current economic crisis (e.g., exchange rate depreciation, high inflation rate of 10% in 2015, weakened macro-economic stability) and political crisis (manifested in the form of low approval ratings and large-scale demonstrations since 2013) have compromised investment activity and consumer confidence (Bertelsmann, 2016; The World Bank, 2016b). Compared to other middle-income countries, the quality of government services in relation to spending also remains poor.

The Emergence and Development of Programme Evaluation in Brazil

The practice of programme evaluation in Latin America evolved in relation to the political environment that prevailed at different points in time. Three different political phases influenced its evolution: (a) the development of a welfare state, prevailing until the 1970s; (b) the emergence of authoritarian governments and the period of neo-conservatism typical of the 1980s and the 1990s; (c) the reintroduction of progressive governments that supported open politics, social mobilisation, and a collaborative relationship between the state and civil society (Neirotti, 2012).

During the welfare state period, data such as educational and economic statistics, were regularly collected and analysed for regulatory and planning purposes. Systematic programme and policy evaluations were not carried out during this period. The emphasis was on feasibility studies and process inspections following project implementation, with little room for formative re-planning (Neirotti, 2012).
With the process of denationalisation, privatisation, deregulation, and decentralisation, characteristic of the second political phase, there was a move towards formal evaluation of policies. This period saw the formation of structures for the systematic evaluation of state services, such as the National System for Results Evaluation of the Public Administration in Columbia, the National System of Evaluation in Costa Rica, and evaluation mechanisms built into the Brazilian Action Plans at federal, state and municipal levels (Neirotti, 2012). During this period, evaluation focused on the monitoring of outsourced state functions and the performance of the public administration.

Following the increase in political mobilisation, the development of civil society organizations, and the processes of state modernisation, which characterised the third political phase, there was a growing need for the professionalisation of the field and training of evaluators. Support for evaluation research and professionalisation of the field was provided by the government and international organizations (Neirotti, 2012).

Movement through these three phases has resulted in a shift in terms of the focus and purpose of evaluation in Latin America, from a social planning orientation to a

results-driven orientation. Table 8 presents the characteristics of the evaluation function within the three different political stages outlined in this section.

Table 8

Characteristics of Evaluation within Different Political Phases

	State-centric Phase 1	Neoconservative Phase 2	State/Society Articulation Phase 3
State presence	Welfare	Withdrawal of the state	Return of the state
Evaluation system orientation	Production of statistical data to support planning	Systematic research with an emphasis validity information	Focus placed on valid knowledge, communication and use of knowledge
Evaluation focus	Ex ante evaluation Audits and process control	Ex post evaluation Results	Results in relation to processes
Methodologies	Statistical analysis Feasibility analysis	Quantitative (impact analysis)	Quantitative/Quantitative combination

Note. Table adapted from Neirotti (2012).

Given that Brazil's political landscape can be captured in the three distinct phases outlined above, one can reasonably assume that the development of the evaluation function in the country followed more or less the same pattern summarised in Table 8. The next section outlines specific developments, that illustrate the evolution of the field in Brazil, following the re-introduction of a democratic government in 1989 (phase 3).

Brazil's re-democratisation process has redefined the role and participation of citizens in the decision making process, and enhanced the collaboration between the private and public sector in policy implementation, through joint ventures. Both of these developments have increased the need for transparency, accountability and stronger M&E systems (Henriques, Pinho, Azevedo, & Newman, 2010).

Evaluation in the public sector emerged in Brazil in the 1990s, with the development of the Master Plan of the State Apparatus Reform, issued in September 1995 (Henriques, et al., 2010). This new emphasis on results-oriented management not only influenced the work of the Brazilian federal government but also the activities of international and civil society organizations. The Ministry of Planning, Budgeting, and Management (MPOG) has been the main driver in the process of incorporating a results-oriented focus into sector-oriented public policies, and developing an evaluation culture within the public administration system to enhance policy formulation and implementation (Chianca, 2007; Henriques, et al., 2010). The effort started with the implementation of the 2000-03 Pluri-Annual Plan (PPA) and associated innovations, such as: (a) the development of implementation and outcome indicators, which promoted the emergence of a results-oriented management culture and concepts of M&E; (b) the assignment of programme managers, trained in the use of project/programme management tools; and (c) the publication of a methodological guide to programme indicators, reporting on programme progress (Henriques et al., 2010). With the implementation of the PPA, as a key vehicle for addressing major government priorities, the main emphasis was on programme objectives and performance indicators and the linkage between evaluation results and allocated budget (Burdescu, del Villar, Mackay, Rojas, & Saavedra, 2005).

Sector-oriented ministries, in turn, gradually began to incorporate M&E policies in the fields of education, health, and social development. The Ministry of Social Development and Hunger (MDS), created in 2003, was the first ministry to establish a specialised bureau for information management. The bureau is responsible for baseline research, beneficiary surveys, and impact assessments to support policies on, for example, cash transfers, social assistance, and food security (Henriques, et al., 2010).

Selected states and municipalities have shown increasing interest in the M&E agenda. The state of Minas Gerais, for instance, introduced results-based management in 2003, a robust M&E system in 2007, and developed innovative practices that emphasised the use of indicators to direct public administration (Guimaraes, & Campos, 2010).

Key Players in the Brazilian Public Service Evaluation

Among Brazil's state institutions, the contribution of the State System for Data Analysis Foundation (SEADE) is also worth noting. In the last decade, the Foundation has focused on the creation and strengthening of methodologies for the formulation, monitoring, and evaluation of public policies and programmes, and providing support to agencies that deliver or coordinate government activities. Similarly, Brazilian development banks, such as the Banco do Nordeste do Brasil, and the Brazilian private sector have performed a key role in strengthening the M&E agenda. Some initiatives include the Crediamigo programme evaluation, and the assessment of the Northeast Financing Constitutional Fund carried out in partnership with the World Bank, as well as the establishment of the Competitive Brazil Movement in 2001, which further promoted the adoption of results-based management practices (Henriques, et al., 2010).

Finally, the contribution of international organizations to the M&E agenda in Brazil cannot be overlooked. The World Bank for instance, provided the federal government with technical assistance in developing M&E systems and activities, through the implementing of the Brazil Evaluation programme (Brasil Avaliação, BRAVA). The BRAVA was launched in 2005 and concluded in 2009. Specific objectives of the programme included: (a) assisting the federal government and its ministries in the design of results-oriented management systems, (b) promoting the use and dissemination of the information generated by the M&E system, and (c) creating an M&E culture.

Evaluation Capacity-Building in Brazil

The Brazilian M&E network (BMEN).

As the previous section suggests, Brazil has made significant progress in evaluation policy and practice (Henriques et al., 2010). This existing knowledge and expertise is however not fully systematised and disseminated to the desired level. This situation, along with other important challenges such as, a lack of evaluation professionals, limited networking opportunities for evaluators and those interested in the field,

scarce evaluation literature in Portuguese, and a lack of evaluation training courses, led to the creation of the BMEN in 2002 (Firme, Letichevsky, Dannemann, & Stone, 2009). The BMEN was launched as a chapter of the Latin America and the Caribbean Monitoring and Evaluation Network, with support from of the Fundação João Pinheiro (JPF), the Inter-American Development Bank (IDB), and the World Bank.

The four main goals articulated in BMEN's mission statement are to:

- Organize individuals and organizations, directly or indirectly involved with evaluation debate and practice.
- Create, disseminate and manage knowledge on evaluation.
- Promote the training and development of professional evaluators.
- Advocate the inclusion of evaluation practice in management and planning strategies of public and private institutions.

By December 2004, BMEN had secured 264 subscribed members and established regional hubs in six state capitals within the North Eastern, Central and South Eastern regions of Brazil. According to the International Organization for cooperation in Evaluation (IOCE)'s 2012 and 2015 partners profile survey, BEMN reported a total membership of 2645 in April 2012, and 6000 members in March 2015. Affiliates include key stakeholders (individual and institutions) involved in the M&E field in Brazil and abroad (Henriques et al., 2010). Table 9 presents the 2015 member profile and distribution of BMEN as per IOCE's records.

Table 9

Reported BMEN Membership as at March 2015

Member profile	*n*
Government	2550
NGOs	450
Academics	600
Private sector consultants	620
Other	1780

In order to promote the interchange of experiences and knowledge, and strengthen the culture of M&E in Brazil, BMEN has held four national seminars in the period 2008-2012 (including one-day workshops and short courses), with an average of 300 participants. The national seminars organised by BEMN focus on two main aspects: the design and institutional strengthening of M&E systems for public policies and programmes, and M&E methodologies and practices (Henriques et al., 2010).

One of BMEN key achievements is the publication of the Brazilian M&E Journal, in partnership with the Ministry of Social Development. In 2009, BMEN also created a platform for exchanging M&E knowledge and experiences over the internet. This platform includes expert blogs on M&E, videos of lectures and seminars, a calendar of international events and national publications, and discussion forums.

University-based evaluation training programmes and other training options in Latin American and Caribbean Countries (LAC).

Documented evaluation training options currently available in Brazil, beyond those provided by BEMN, are scarce. The only documented reference is the certification course on Evaluation of Social Programmes, offered by Brazilian Ministry of Social Development, the National School of Public Administration and the National School of Public Health in association with the Institute of Social Studies (The Netherlands) in 2005 (Chianca, 2007).

To get a sense of the different training options available to Brazilian evaluators it is important to consider the availability of courses offered across LAC. A number of on-line evaluation courses have been offered by the following institutions in/since 2005:

- The University Nacional del Litoral and the Technological Technical Assistance Centre for Public Organizations (TOP), in Argentina, offer an on-line certification course in Outcome and Impact Evaluation of Public Organizations and Programs.
- The Inter-American Development Bank (BID) offers a series of on-line courses including: logic models for project design, project M&E, evaluation of environment impact, and institutional analysis.

Other institutions in LAC offering specialised courses in evaluation include:

- The Center for Studies in Economic Development (CEDE) at the University of Los Andes in Bogotá, in Colombia.
- The Latin American Institute for Social and Economic Planning (ILPES).
- The Argentinean Evaluation Association (AAE).

Current Challenges in Evaluation in Brazil

The training of professionals for the practice of evaluation, as well as developing evaluation capacity among programme stakeholders, still remain a challenge in Brazil (Firme et al., 2009). As a result, evaluations of varying quality are currently being produced, including those that do not have a guiding evaluation policy or are lacking in terms of transparency. In addition, there is considerable debate among evaluators and evaluation stakeholders around a number of issues, including: (a) what comprises quality evaluations, (b) the importance of constructing an evaluation culture, and (c) the extent to which evaluations have contributed to the growth of stakeholder's social capital. In order words, evaluation culture (i.e., a shared understanding and acceptance amongst stakeholders of the need and practice of evaluation) is not fully developed in Brazil. Evaluation policy (i.e., a set of guidelines that establishes rules, procedures, and standards to properly plan, implement and

use evaluation) is also still not well-documented. Conducting evaluation in the absence of evaluation policy and culture compromises the quality of evaluations and limits evaluation use (Firme et al., 2009).

Summary

This sub-section provided an overview of how evaluation progressed in Brazil- a relatively new democracy and emerging economy. As highlighted in this sub-section, M&E emerged in Brazil only in the 1990s. The effort was largely state-initiated. To date, evaluation is not yet firmly institutionalised within the democracy. Training options available to M&E professionals are still limited and the evaluation community in Brazil is still grappling with issues of evaluation standards. On the whole, both evaluation culture and capacity is relatively weak in Brazil.

SOUTH AFRICA

South Africa's Key Historical and Political Markers

South Africa's 1994 transition from apartheid to a constitutional democracy remains one of the most striking political transitions in history (The World Bank, 2016d). During the apartheid regime, racial discrimination was institutionalised with the enactment of apartheid laws in 1948. These laws governed every aspect of social life, from housing to healthcare, and placed a number of restrictions on the South African population. In 1950, the Population Registration Act required all South Africans to be classified into one of four racial categories: White, Black (African), Coloured, or Indian Classification into these categories was based solely on appearance, social acceptance, and descent (Kalley, Schoeman, Andor, 1999; Worden, 2011).

In most respects, apartheid was a continuation of the segregationist approach of previous governments. The system was however more systematic. The apartheid policy, termed as separate development, was designed to mask the discriminatory nature of official policy making. Black, Coloured and Indian people were subject to widespread discrimination at all levels.

After a long negotiation process, sustained violence from the right wing, and support from the international community, South Africa's first democratic election was held in April 1994 under an interim Constitution. Since 1994, the African National Congress (ANC) has won all national democratic elections.

The new democratic constitution called for democratisation, socio-economic upliftment, a culture of human rights, and improved service delivery, with a commitment to improve the lives of all South Africans, particularly the poor (Seo, 2008). A significant milestone in the democratisation of South Africa includes the country's ability to host subsequent elections in a peaceful and fair manner, and with high levels of participation. Since the democratic elections in 1994, one key priority of the government has been to integrate the country into the global political, economic and social system. South Africa is increasingly gaining prominence on the international stage and has become an active participant in events such as the Annual Meetings of the International Monetary Fund and the World Bank, the G-20, and the G-24. The country also joined the BRICS nations in April 2011 (The World Bank, 2016d).

South Africa's Current Socio-Economic Standing

South Africa is an upper middle-income country with a gross domestic product (GDP) of US $312.789 billion in 2015 (The World Bank, 2016d). The country exhibits many characteristics associated with developing countries, including an uneven distribution of wealth and income. It has however one of the most progressive constitutions in the world and a stable political environment. South Africa has maintained a steady GDP growth rate up to a decade after the global financial shock of 2008-2009 due to its macroeconomic stability (The World Bank, 2016e).

Since democratisation in 1994, the South Africa Government's development agenda has focused on poverty alleviation and the provision of social services to historically disadvantaged groups (Patel, 2008; Seo, 2008). A number of poverty reduction strategies have been implemented. These include the Reconstruction and Development Programme (RDP) in 1994, the Growth, Employment and Redistribution (GEAR) programme in 1996 and the Accelerated and Shared Growth

Initiative for South Africa (ASGISA) in 2006. Over the past decade, South Africa has made considerable progress in the areas of education, housing, healthcare, water and sanitation, and social security, following substantial increases in the social sector budget. There has been, for instance, an unprecedented increase in the number of South Africans receiving social grants in past five years (approximately 15 million in 2014), with the majority of recipients being parents or carers receiving the Child Support Grant (South African Social Security Agency, 2014). Similarly, the government has made substantial investments in public education in order to improve access to and quality of education at all levels.

South Africa's Current Socio-Economic Challenges

Despite sustained improvements across a number of sectors, South Africa continues to face a number of development challenges relating to policy implementation and the delivery of services (The World Bank, 2016e). The country's development challenges are multidimensional, and deeply entrenched.

There is still a high incidence of absolute and relative poverty, and income inequality. The country's distribution of income and wealth is among one of the most unequal in the world, with a Gini coefficient of 0.65 in 2011 (The World Bank, 2016d). Two factors account for the high level of income inequality in the country. The first factor is the entrenched spatial patterns that perpetuate apartheid's segregation, with a large proportion of the population living in townships and informal settlements. The second factor is the country's inability to generate sufficient jobs. The unemployment rate has increased from 22.7% in 2008 to 25.1% in 2015 (The World Bank, 2016e). The country also has a significant shortage of high-skilled workers.

Despite improved access to basic services such as education, healthcare, and water and sanitation, the quality of service delivery in the country is still weak (The World Bank, 2016d). Education expenditure as a percentage of total government expenditure dropped from 20.6 in 2012 to 19.1 in 2014 (The World Bank, 2016e) and the provision of high quality education is still one of the greatest challenges facing South Africa today. Authors such as Taylor, Muller, and Vinjevold (2003, p.41) have argued that "learners' scores are far below what is expected at all levels of the

schooling system, both in relation to other countries (including other developing countries) and in relation to the expectations of the South African curriculum". South Africa is still outperformed by countries spending less per capita on education.

Other development challenges endemic to South Africa include the high HIV/AIDS infection rate and tuberculosis prevalence rate. Although South Africa spends an estimated 8.8% of GDP on health (as per 2014 World Bank statistics), the country still has poor health indicators and does not compare well with countries with similar or lower national income and health expenditure per capita. The country's current infrastructure, although considered as highly developed by African standards, also continues to lag behind when compared with countries of a similar development level (Bogetić, & Fedderke, 2006).

The Emergence and Development of Programme Evaluation in South Africa

The years leading up to South Africa's first democratic election in 1994 were marked by the unprecedented increase in interventions launched by non-governmental organizations (NGOs). These interventions aimed to address the socio-economic inequalities created by the apartheid regime (Louw, 1995; Mouton, 2010). For many decades, NGOs have received extensive support from international donor agencies. Before 1994, there were few regulations and stipulations attached to donor funding. Funds were channelled directly to NGOs and the highest expectation in that period was the provision of occasional reports and audited financial statements (Podems, Goldman, & Jacob, 2014). The NGO landscape changed drastically after the first democratic election. Donor agencies, in support of the new democracy, started channelling funds primarily through government vehicles and more stringent criteria were attached to support. There was progressively increased pressure on NGOs to demonstrate accountability, with donors frequently commissioning programme evaluation. With funding becoming increasingly difficult to obtain and dependent upon some form of evaluation, NGOs had to adjust their role from voluntary organizations to service providers (Mouton, 2010).

Even though large donor organizations such as USAID South Africa, the Kellogg Foundation, and the European Union conducted large-scale evaluations prior to

1994, the literature suggests that many of these evaluations were not designed and implemented by local NGO staff. Instead, these donors would typically appoint external evaluators to assess the effectiveness of their investment. There was some awareness of the need to evaluate (in the widest sense of the word) among local stakeholders, but systematic programme evaluation only took off from 1994 onwards in South Africa (Louw, 1995; Mouton, 2010). A culture of programme evaluation only began to emerge locally in response to donors' call for increased accountability and use of evaluation tools (such as logic frameworks) and practices. In addition to the external pressure exerted by donors, a variety of factors from within the NGO sector also fuelled the need for greater accountability post 1994 (Mouton, 2010). For example, with the growth of the NGO sector, NGOs gained greater power in influencing policy and lobbying for greater transparency and accountability in the government and private sector. In order to strengthen their position, NGOs were forced to mirror accountability and transparency. It should be noted that programme evaluation was still not common practice at that stage.

Although programme evaluation gained entry into South Africa through the donor community, it should be highlighted that it only firmly established itself when the public sector institutionalised this practice, through the introduction of various mechanisms, strategies and accompanying legislative mandates (Mouton, 2010). Prior to 1994, monitoring and reporting practices that existed within the public sector were geared towards generating information for control purposes (Madzivhandila, 2010). Post-1994, initial efforts to consolidate this information for decision-making and improvement purposes were undertaken. This attempt was in line with the democratic government's agenda to transform the way in which the public sector operated. This transformation process occurred in three distinct phases, namely, the rationalisation and policy development phase (1994-1999), the modernisation and implementation phase (1999-2004) and the accelerated implementation phase (2004-current) (Madzivhandila, 2010). The major developments in M&E within the South African public sector occurred in the third phase.

Prior to 2004, monitoring and evaluation activities were not systematically undertaken in the South African public sector (Cloete, 2009; Madzivhandila, 2010). These activities were undertaken sporadically by government departments for

annual reporting purposes. Some departments were more rigorous than others in the process. The Department of Land Affairs, for instance, adopted M&E quite early on. The department implemented an extensive Geographical Information System to assist project monitoring in 1995 and undertook diagnostic studies similar to programme evaluation (Engela & Ajam, 2010; Mouton, 2010).

Legislation mandating systematic M&E across national and provincial departments only came about in 2004/05 (Madzivhandila, 2010). In line with the White Paper on the Transformation of Public Services and other policy documents advocating M&E, the South African Cabinet adopted a strategy to standardise the way in which M&E is practised throughout the public sector in South Africa (Cloete, 2009; Mouton, 2010). A Government-wide Monitoring and Evaluation System (GWM&ES) was developed as part of this strategy. The following factors motivated the cabinet to develop the GWM&ES:

- The increased importance attached to M&E systems worldwide.
- The need to report back on the United Nations Millennium Development Goals (MDGs).
- The fact that the country hosted the World Summit on Sustainable Development in 2002, in the absence of a national system to monitor sustainable development, thus violating the requirement imposed by the Rio Convention of 1992.

The conceptualisation of the GWM&ES was a key milestone for good governance in South Africa. The GWM&ES served as a vehicle to formalise and streamline monitoring and evaluation activities in government, manage performance and measure service delivery of government departments, and improve public service delivery (Madzivhandila, 2010; Rabie, 2010). The GWM&ES was envisaged to culminate into:

- enhanced quality of performance information and analysis within departments and municipalities;

- improved monitoring, evaluation, and reporting of outcomes across the government through, for example, the Government Programme of Action bi-monthly Report, and the Annual Country Progress Report based on the national indicators;
- improved monitoring and evaluation of provincial outcomes against Provincial Growth and Development Plans;
- projects to improve M&E performance in selected institutions across government; and
- M&E capacity building initiatives to foster a culture of governance and decision-making which responds to M&E findings.

The initial proposal for the GWM&ES was revised and updated in 2007 (Cloete, 2009). A number of guiding frameworks were published by key players, such as The Presidency, The National Treasury, and Statistics South Africa, to support the implementation of the GWM&ES (Madzivhandila, 2010). These include: (a) a Policy Framework for the Government-Wide Monitoring and Evaluation System (The Presidency), (b) a Framework for Managing Program Performance Information (The National Treasury), and (c) a South African Statistical Quality Assessment Framework (Statistics South Africa). The development of evaluation frameworks and systems were entrusted to each department. Even though progress was slow, the GWM&ES has been gradually institionalised at both national and provincial levels (Abrahams, 2015; Cloete, 2009; Madzivhandila, 2010).

Other parallel developments in the public sector include the establishment of a dedicated Performance Monitoring and Evaluation Department (later referred to as the Department for Planning, Monitoring and Evaluation) within the Presidency in 2010, and the appointment of a Minister of Performance Monitoring and Evaluation (Podems et al., 2014). The creation of a dedicated department of Performance Monitoring and Evaluation (DPME) is indicative of a strengthened commitment towards M&E in the South African public sector (Engela & Ajam, 2010). Prior to 2011 there was no standardisation of evaluation in government. The emphasis was on monitoring. It is only in 2011 that the DPME's role extended to evaluation. The DPME developed a National Evaluation Policy Framework (NEPF), which was

approved by Cabinet in November 2011 (Podems et al., 2014). The NEP reflects high-priority government interventions (i.e., those that relate closely to priority outcomes, involve high government spending, and address issues of considerable public interest) and is updated annually (DPME, 2016). The NEPF “has attempted to shift government from a compliance culture to one that has a greater emphasis on improvement, learning, and efficiency” (Podems et al., 2014, p.75).

Key Players in South Africa’s Public Sector Evaluation

A number of constituents within the government are involved in Public Sector Evaluation. A brief overview of the roles and contribution of some of the key players are presented in the next section.

Department of Planning, Monitoring and Evaluation (DPME).

The DPME was responsible for coordinating the implementation of the GWM&ES (the NEPF’s forerunner) and reviewing the data architecture of government departments so that the required performance information is generated. Amongst other initiatives, the DPME established a Government-Wide Monitoring and Evaluation Learning Network (GWM&ELN) in order to facilitate the development of evaluation capacity within the government (Madzivhandila, 2010), and produced standards for evaluation in Government in 2012 (in partnership with the South African Monitoring and Evaluation Association). Recent notable milestones of the DPME include the development of 21 guidelines and templates on different components of the evaluation process to support departments conducting evaluations. In 2015, DPME produced a guideline on how to develop Departmental Evaluation Plans (DPME, 2016). The DPME currently has 141 evaluations logged on its evaluation repository, publicly accessible on its website. DPME has trained over 1200 staff involved NEPF evaluations. The DPME’s custodial role for M&E is similar to the functions of National Treasury for financial management (Abrahams, 2015).

The National Treasury.

The National Treasury supports the Ministry of Finance in determining fiscal policies. As such, its responsibilities involve the monitoring of economic indicators. In addition, the National Treasury measures the attainment of objectives and targets set in the Estimates of Expenditure, and evaluates whether public expenditure has achieved value for money. These evaluations are published in key documents such as the Budget Review, the Provincial Budgets and Expenditure Review, and the Local Government Budgets and Expenditure Review. The Treasury's involvement in the GWM&ES revolved around ensuring that information on inputs, activities, outputs and outcomes inform budgetary planning, and accountability reporting, and expenditure control (Madzivhandila, 2010).

Office of the Public Service Commission (PSC).

The PSC has the constitutionally prescribed function to monitor and evaluate the organization and administration of the Public Service and propose measures to improve its performance (Madzivhandila, 2010). Their mandate also includes the evaluation of government programmes.

Evaluation Capacity-Building in South Africa

There were few independent practitioners/consultants/researchers conducting evaluations as a full-time activity in the 1990s (Louw, 1995). Most practitioners were affiliated to university departments (e.g., Sociology, Education, and Psychology departments) or to health and educational policy analysis units such as the Human Sciences Research Council (HSRC) and the Medical Research Council (MRC). In addition, those involved in evaluation often had to rely on their research methods training in their primary discipline as both formal and informal training opportunities in evaluation methods were non-existent in South Africa in the early 1990s. Another difficulty faced by early evaluators was that they had to work in isolation since there was no formal network or association of evaluators. Mouton (2010) termed this first cohort of researchers involved in programme evaluation as first generation evaluators. Through in-depth interviews with selected first generation experts in the

field, Mouton (2010) established that their programme evaluation knowledge (evaluation theory, design, and methodology) was mainly self-taught. These evaluators had to rely extensively on their own understanding of what programme evaluation entailed and adapt their methodologies through application and practice. These individuals resorted to a number of strategies to build their M&E capacity and establish themselves in the field. These included: (a) doing extensive reading on the field, (b) attending international conferences, (c) utilising learning aids from development organizations such as the World Bank, (d) and liaising with international experts.

Few formal training opportunities were available in the early 1990s. There has however been an expansion, both in terms of the number and the depth of formal programme evaluation courses offered by capacity-building organizations and Higher Education institutions, in recent years.

The South African Evaluation Association (SAMEA) and the African Evaluation Association (AFREA).

Different voluntary associations such as SAMEA and AFREA have also been active in evaluation capacity-building and transforming the field into a more organized profession. SAMEA was founded in November 2005, with the aim of formalising the South African Evaluation Network (an informal network of evaluators) and cultivating a vibrant community that will strengthen the development of M&E as an important discipline, profession, and accountability instrument in South Africa. SAMEA offers a number of informal training opportunities to evaluation practitioners and a platform for consultancies to advertise their professional development programmes. Activities undertaken by SAMEA include: (a) hosting biennial conferences and regular seminar series, (b) maintaining an online resource database for evaluation practitioners, (c) round table discussions of topical issues in M&E, and (d) updating a repository of member evaluators (SAMEA, 2016).

A set of evaluation guidelines was established by the AFREA to assist African evaluators in planning evaluations, negotiating evaluation contracts, and ensuring adequate completion of a given evaluation. A tentative set of guidelines emerged

following a review of U.S. Program Evaluation Standards, undertaken in a series of workshops and meetings facilitated by evaluators in Africa. These tentative guidelines were presented to a plenary session at the Inaugural Conference of the AFREA in September 1999 (AFREA, 2002). The recommendation was to have these guidelines reviewed by national evaluation associations and networks in Africa and pilot them in several countries. Eleven national and regional networks and associations suggested modifications and endorsed the final version of the guidelines in 2002. The finalised set of African Evaluation Guidelines is framed around the following four categories (AFREA, 2002):

- Utility–the utility guidelines are intended to help to ensure that an evaluation will serve the information needs of intended users and be owned by stakeholders.
- Feasibility–the feasibility guidelines are intended to help to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.
- Propriety–the propriety guidelines are intended to help to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.
- Accuracy–the accuracy guidelines are intended to help to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated.

SAMEA does not have their own set of evaluation standards but instead support the African Evaluation Guidelines.

University-based evaluation training programmes in South Africa.

Programme evaluation has evolved from being an isolated module within a bigger academic programme to a fully-fledged programme on its own. While no South African university trains entry-level evaluators through an undergraduate qualification, a number of universities offer post-graduate qualifications in evaluation,

in the form of short courses and/or master and doctoral programmes (Madzivhandila, 2010).

The Centre for Research on Evaluation, Science and Technology (CREST) at Stellenbosch University offers three post-graduate programmes in monitoring and evaluation: (a) a postgraduate diploma in Monitoring and Evaluation, (b) an MPhil in Monitoring and Programme Evaluation, and (c) a PhD in Evaluation Studies. The MPhil and PhD programmes have been offered for the first time in 2012. The University of Cape Town presents two postgraduate courses in programme evaluation since 2007: (a) an MPhil in Monitoring and Programme Evaluation, and (b) a PhD in Programme Evaluation. The Department of Public Governance at the University of Johannesburg offers a Master's degree in Policy Evaluation, while the School of Health Systems and Public Health at the University of Pretoria offers a Master of Public Health (MPH) degree programme with an M&E concentration, in collaboration with MEASURE Evaluation, since 2012. MEASURE Evaluation is a team of experienced organizations, funded by USAID, to support improvements in monitoring and evaluation in population, health and nutrition worldwide. Other university-based programmes include the postgraduate diploma in Monitoring and Evaluation in the Public Sector, offered by the Department of Public Administration at the University of Fort Hare, and the postgraduate diploma in Public and Development Sector Monitoring and Evaluation at the University of Witwatersrand. The different university-based evaluation programmes currently available in South Africa are summarised in Table 10.

Table 10

University-Based Study Programmes in Evaluation in South Africa

University	Programme Type
The Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University	Postgraduate diploma in Monitoring and Evaluation MPhil in Monitoring and Programme Evaluation PhD in Evaluation Studies
Department of Management Studies, University of Cape Town	MPhil in Monitoring and Programme Evaluation PhD in Programme Evaluation.
Department of Public Governance, University of Johannesburg	Master's degree in Policy Evaluation
The School of Health Systems and Public Health, University of Pretoria; MEASURE	Master's degree in Public Health (MPH), with an M&E concentration
Department of Public Administration, University of Fort Hare	Postgraduate diploma in Monitoring and Evaluation in the Public Sector
School of Governance, University of Witwatersrand	Postgraduate diploma in Public Sector Monitoring and Evaluation Master's degree in Public Sector Monitoring and Evaluation

It is clear from Table 10 that university-based evaluation training programmes in South Africa are only offered at a post-graduate level. It should be noted that certification courses in Monitoring and Evaluation are also offered by the Institute for Monitoring and Evaluation (IME) at the University of Cape Town and the University of Johannesburg in collaboration with the Center for Learning and Results on Evaluation (CLEAR-Africa) and the World Bank. The newly established National School of Government, on the other hand, offers introductory and credit-bearing courses in Monitoring and Evaluation for the public sector (Abrahams, 2015).

Other evaluation capacity-building organizations in South Africa.

A number of consultancies, such as InsideOut, Southern Hemisphere, and Impact Consulting, provide customised in-house training in M&E and offer a number of M&E related services. Donor agencies such as the Carnegie Foundation, Centre for Aids Development Research and Evaluation, the United States President's Emergency Plan for Aids Relief in HIV/AIDS issues, and the World Bank have also assisted in advancing the field of M&E and building evaluation capacity in South Africa (Madzivhandila, 2010). These donor agencies actively support national, provincial and local programmes by providing technical assistance and training.

Current Challenges in Evaluation in South Africa

It is difficult to draw conclusions about the actual status of programme evaluation in South Africa (Mouton, 2010). First, the history of evaluation is too brief, with current efforts devoted to developing and implementing policies and procedures. Second, there has been no assessment of how knowledge is constructed and used practice. Third, there are no systematic studies investigating the influence of evaluation on policy and decision-making (Madzivhandila, 2010). Fourth, there are no clear patterns on how M&E systems are developing nationally and within the provinces. While the recent developments within the public sector and beyond signal the increased awareness of the need to improve systems and practices, and develop evaluation capacity and capability, and to conform to evaluation standards in general, Madzivhandila (2010) and Cloete (200) argue that the progress in developing M&E systems and the current rollout of capacity-building initiatives have been slow. For instance, only a few provincial departments have established M&E units that are fully capacitated in terms budget, staff and systems. The fundamental issues delaying this process, as identified by Cloete (2009) and Madzivhandila (2010), include: (a) absence of a strong evaluation culture within the public service, (b) absence of information systems enabling departments to provide useful performance information, and (c) the lack of an agreed framework of performance criteria and of other deficiencies in the formal reporting requirements. In addition, as remarked by Abraham (2015), there is a gap between policy formulation and implementation within the public sector, with the DPME using a top-down approach.

This type of approach precludes inclusivity and scope for implementing agents to interrogate the policy.

Evaluation capacity within the public sector is also weak. Podems et al. (2014) discussed the lack of evaluation expertise in the public sector by drawing on the findings of a 2013 DPME study. According to this study, only 42% of provisional departments and 52% of national departments were confident in managing evaluations.

Despite these challenges in the public sector, there are a number of factors that might facilitate the advancement of the field of programme evaluation in South Africa. These include: (a) an upsurge in formal and information M&E training options; (b) the advantage of late coming learning from other countries' experiences and international evaluation best practice; (c) context relevant academic publications, such as the Evaluation *Management in South Africa and Africa* text edited by Cloete, Rabie and De Coning (2014); and (d) preparedness to enhance evaluation systems and practices (Abrahams, 2015; Madzivhandila, 2010). In addition, there is a steady increase in the number of practitioners having an interest in the evaluation profession, as evidenced by the growing number of active SAMEA members (from 121 active members in 2007/08, 204 in 2008/09, to 348 in 2009/10). It is difficult to establish the actual size of the current evaluator workforce in South Africa as no study has been conducted in this area. The SAMEA directory is at this stage the most up to date resource on the current evaluator workforce. SAMEA currently has 401 active members: 36% in government, 31% in the private sector, 11% in NGOs, and 8% in academic institutions (Abrahams, 2015).

Summary

This sub-section provided an overview of how evaluation progressed in South Africa- a relatively new democracy and emerging economy. As highlighted in this sub-section, M&E gained entry into South Africa through the donor community in the 1990s, but firmly established itself when the public sector institutionalised this practice across national and provincial departments in 2004/05. The conceptualisation and implementation of the GWM&ES was a key milestone in

operationalising the government's mandate to improve service delivery, and in advancing the M&E agenda. As discussed, South Africa has made significant progress in terms of developing evaluation policy and standards, but still lags behind with regard to the implementation of M&E systems. Evaluation culture and capacity is also not fully developed despite the upsurge in formal evaluation training options and increased preparedness to enhance evaluation systems and practices in recent years.

THE UNITED STATES OF AMERICA (USA)

USA's Key Historical and Political Markers

The USA is a constitution-based federal nation with a strong democratic tradition. The constitution, drafted in 1789, established a federal system which remained unchanged since its inception. The American constitution is one of the world's oldest written constitutions and has served as a model for a number of other constitutions around the world. The federal political structure comprises 50 states.

The USA is the world's leading economic power, with a GDP of US $17.947 trillion in 2015 (The World Bank, 2016). With the global economic downturn in 2008, the country's GDP contracted until the third quarter of 2009, making this longest and most challenging economic recession since the Great Depression of the 1930s.The economic recovery has gained momentum since the first half of 2011.

The USA's Current Socio-Economic Standing

The United States performs well in terms of overall socio-economic indicators. The country has the highest average household net adjusted disposable income per capita, household net financial wealth and average earnings among 34 Organization for Economic Co-operation and Development (OECD) countries. Educational attainment is also among the highest, with 89.6% of the adult working-age population having completed at least an upper secondary education (OECD, 2016a). In 2015, 44% of the working-age population had attained a tertiary degree, which is considerably higher than the OECD average of 33% (OECD, 2015). The

unemployment rate of 4.7% is well below the OECD rate of 6.5% (OECD, 2016b). There has also been a sharp decline in long-term unemployment in 2016.

The USA has a robust welfare state, with an average welfare spending of 19% of GDP over the period 2011- 2016 (OECD, 2016c). In 2014, total social spending in the USA was the second highest in the world (OECD, 2014). Welfare spending on health care rose from 3.5% of GDP in 1980 to 8% in 2012. Although health spending has declined considerably in recent years, it is still higher (on a per capita basis) than in all other 34 OECD countries (OECD, 2015b).

USA's Current Socio-Economic Challenges

Despite being the world's richest economy and having a robust welfare state, the USA still faces a range of socio-economic challenges. The country has the second highest rate of homicides amongst all OECD countries. Life expectancy is lower than most OCED countries (OECD, 2016a). This state of affairs has been attributed to the relatively poor health-related behaviours of Americans and the highly fragmented nature of the USA health system (OECD, 2015b). Child mortality is also among the highest in the OECD. In fact, child outcomes are poor in general, with 21.1% of children reporting poor health (OECD, 2016a). Child income poverty is significantly higher than the OECD average, with 20.5% of children living in a household with a disposable income of less than half of the American median income. There are stark regional inequalities in income. For example, the average household adjusted disposable income in the District of Columbia is more than double that in Idaho. In fact, the USA has the widest income disparities among OECD countries. There are also regional differences in the rate of unemployment (e.g., 2.9% in North Dakota compared to 7.9% in the District of Columbia). This 5% gap is larger than those observed by other OECD countries like Australia. Fourteen point four percent (14.4%) of young people are neither employed nor in education or training ("NEETs"). This percentage is substantially higher than in countries like Germany and Japan (OECD, 2016b). NEETs, especially low-skilled NEETs, are at risk of being left permanently behind in the labour market.

The Emergence and Development of Evaluation in the USA

The first strand of evaluative inquiry emerged in the USA as early as the 1900s, in the form of agricultural research (Chelimsky, 2006). This type of research applied experimental designs and statistical analyses to isolate agricultural practices that would lead to the largest crop yields. A second evaluative strand began in the 1950s, with efforts to justify resource allocation and management of defence programmes. The Rand Corporation can be credited for the initial development of this evaluative strand, which eventually grew into the Planning, Programming and Budgeting System (PPBS) of the Department of Defence. Developed largely by economists and political scientists, the system used techniques such as policy analysis, cost-benefit, cost effectiveness and system analysis (Chelimsky, 2006). Contemporary evaluation of educational and social programmes, on the other hand, emerged in the 1960s, following the large-scale funding of such programmes under the banner of the War on Poverty by Presidents Kennedy and Johnson (Chelimsky, 2006; Rist, 1989; Rist & Paliokas, 2002; Stame, 2003; Worthen, 1994). Early in that decade, the U.S. Congress supported the evaluation of such programmes through federal legislation. The Elementary and Secondary Education Act (ESEA) of 1965 for example firmly mandated the practice of evaluation. The ESEA not only provided large-scale funding for compensatory education for disadvantaged youth or innovative educational projects but required recipients of grants to evaluate the outcomes tied to their expenditure. This requirement was initially met with difficulty and minimal success as recipients were not trained in the complex task of isolating the effects of a given programme (Worthen, 1994).

By the late 1960s, federal funds were allocated for the evaluation of social programmes in areas as diverse as vocational rehabilitation, child health, and community initiatives (Leeuw, 2011; Worthen, 1994). These evaluations relied heavily on the methods of applied social science. The randomized controlled trial became the ruling paradigm for evaluation research in the 1960s and the 1970s (Leeuw, 2011).

In the absence of its own evaluation capability, the U.S. Congress relied on the findings and reports of the executive branch and on the U.S. General Accounting

Office (GAO) for independent assessment of the effects of large investments in social programmes (Rist & Paliokas, 2002). Established in 1921 as an independent auditing agency, following the implementation of the Budget and Accounting Act, the GAO was responsible for the investigation of all matters relating to the receipt, disbursement and application of public funds (Rourke, 1978). In 1967, the agency was required by amendments of the Economic Opportunity Act to conduct independent evaluations of anti-poverty programmes and report on the effectiveness of government spending (Leeuw, 2011). By 1969, it was the focal point for evaluation in the legislative branch, producing almost 50 evaluation reports within a two-year timeframe (Melkers & Rossner, 1997; Rist & Paliokas, 2002). Nearly 70% of all professional staff in the GAO was formally classified as evaluators (Rist, 1989).

The increase in federal demand for evaluations, along with key appointments in the GAO, prompted the establishment of a separate Institute for Programme Evaluation in 1980, later renamed the Programme Evaluation and Methodology Division (PEMD) in 1983 (Rist & Paliokas, 2002). The PEMD was staffed by highly-trained social scientists. The PEMD's work was unique among federal evaluation units because its work covered nearly every policy arena, from agriculture to defence, health to transportation, environment to welfare (Grasso, 1996). The division gained international recognition and influenced governments, think tanks and research centres beyond the USA (Melkers & Rossner, 1997).

The upsurge in evaluation activities during the period 1969-1980 can be attributed to the legislative framework that mandated the practice of programme evaluation. Examples of specific laws that underlined the practice of programme evaluation in that period include:

- The Congressional Budget and Impoundment Control Act of 1974 mandating the GAO to develop and disseminate programme evaluation methods for the federal government (Grasso,1996).
- The Legislative Reorganization Act of 1970 requiring the evaluation of programme effectiveness and the development of staff evaluation capacity (Leeuw, 2011).

The strongest legitimisation of programme and policy evaluation ocurred in 1979 when the Office of Management and Budget (OMB) issued circular No. A-117, entitled Management Improvement and the Use of Evaluation in the Executive Branch (Rist & Paliokas, 2002). This circular, which constituted formal policy throughout the executive branch, stated explicitly that all agencies will be assessing the effectiveness of programmes that they implement.

The thriving years of evaluation, however, began to stagnate when Reagan was elected president a year later (Rist & Paliokas, 2002). Programme evaluations began to decline in numbers and scope, following severe cuts in welfare spending and growing fiscal crises experienced during Reagan's Republican administration. Federal evaluation mandates were relaxed, and in some instances abolished (Worthen, 1994). By 1982, government monitoring of federal funding declined drastically and the receding support for programme evaluation was evident. As highlighted in the GAO 1987 report to Congress (Rist & Paliokas, 2002):

- The total number of evaluation units in non-defence departments and independent agencies had declined by 32% between 1980 and 1984.
- The number of professional evaluation staff declined by 22% from 1507 to 1179, between 1980 and 1984.
- The number of evaluation reports declined by 23% from 2114 to 1619 in the same period.

What remained during this period were non-technical in-house evaluations that aimed at internal management rather than evaluations that focused on broader policy questions of overall programme utility and impact (Rist & Paliokas, 2002). There is some evidence that evaluation agencies that did not depend on federal funding, experienced a modest increase in the number of evaluations commissioned to guide programme implementation (Worthen, 1994)

The interest in programme evaluation was revived and there was a resurgence of evaluation activities at both the state and federal level, when Clinton became president in January of 1993 (Melkers & Roessner, 1997). Clinton institutionalised,

through the implementation of the National Performance Review (NPR) strategy, a new focus on performance-based management and accountability that emphasised the shift from inputs and processes to results (Rist & Paliokas, 2002). This coincided with the advent of the New Public Management (NPM) principles in the early 1990s which reconceptualised of the role of the state, and favoured a results-based approach to management (Stame, 2003).

The U.S. congress ratified elements of the NPR by introducing the Government Performance and Results Act (GPRA) in 1993. The GPRA formalised performance measurement and reporting in the federal government and advocated for the incorporation of evaluation results into strategic plans (Rist & Paliokas, 2002). Executive agencies were required to report periodically on their results in achieving their agency and programmatic goals and mandated to prepare performance-based strategic plans. This new legislative structure ensured that government performance data were systematically incorporated into the budget process by driving resource allocation decisions. While the NPR and the GPRA sparked renewed emphasis on evaluation in principle, the initial levels of participation of the fourteen most active evaluation offices in the executive branch in the GPRA and NPR implementation was however lower than expected, as per a study conducted by Wargo (1995). With the concurrent implementation of the NPR and the GPRA, two contradictory forces were at work (Rist & Paliokas, 2002), thus explaining the results of the study. The NPR, whose goal was to create “a government that works better and costs less”, ran head on against the GPRA mandate as it resulted in evaluation offices with fewer resources, evaluation capability and expertise. For instance, a four-year hiring freeze, accompanying attrition, budget constraints, and the retirement of the PEMD leader led to the termination of PEMD in 1996 after 16 years of operation (Grasso, 1996; Rist & Paliokas, 2002). The evaluation function of the GAO was severely compromised with the termination of the PEMD. Only a small residual evaluation unit of seven or eight persons remained within one of the five divisions of the GAO (Rist & Paliokas, 2002).

In 2002, during President Bush’s first administration, the Office of Management and Budget (OMB) developed the Program Assessment Rating Tool (PART) to encourage greater and more consistent use of program performance information in

federal decision-making. Although this tool helped improve the availability of better performance measures, the OMB and the GAO noted that this did not result in their greater use of this information by the Congress and had little influence on resource allocation (Joyce, 2011). The PART was abolished during President Obama's administration. In October 2009, the OMB proposed a three-fold initiative to strengthen federal programme evaluation. The initiative consisted of: (a) regular online updates on all agencies' planned and ongoing impact evaluations, (b) establishing an inter-agency group to promote the sharing of evaluation expertise, and (c) funding selected impact evaluations and capacity strengthening efforts (Office of Management and Budget, 2009). The scope of this initiative can be gauged by the amount of funding allocated to its implementation. For instance, during the 2011 fiscal year the OMB allocated approximately $100 million to facilitate the implementation of 35 rigorous programme evaluations and evaluation capacity-building proposals (U.S. Government Accountability Office, 2011).

High priority performance goals were also established during Obama's administration, as part of performance-oriented reforms (Joyce, 2011). These performance goals are agency directed and serve to facilitate performance monitoring. Examples of high-priority performance goals in the 2011 fiscal year budget are presented in Table 11.

Table 11

Examples of High-priority Performance Goals

Department	Performance Goal	Performance Target
Department of Education	Improve the quality of teaching and learning	Increasing by 200,000 the number of teachers for low income and minority students who are being recruited or retained to teach in hard-to-staff subjects and schools in systems with rigorous processes for determining teacher effectiveness.
Department of Homeland Security	Improve the efficiency of the process to detain and remove illegal immigrants from the United States	Decrease the number of days spent in custody by illegal immigrants before they are removed from the United States from 43 to 41 days in 2010.
Social Security Administration	Improve SSA's Customers' Service Experience on the telephone, in field offices, and online	Achieve an average speed of answer of 264 seconds by the national 800-number.

Note. Table adapted from Joyce (2010)

Finally, with the introduction of the GPRA Modernization Act (GPRAMA) of 2010, Congress further reinforced the mandate to include a discussion of programme evaluations in the strategic plans of agencies (U.S. Government Accountability Office, 2011).

Key Players Currently Involved in programme Evaluation in the USA

It should be noted that there is no central agency for measurement and evaluation in the executive branch (Joyce, 2011). Programme evaluation is decentralised to departments, with the Department of Education, the Department of Housing and Urban Development, and the Department of Health and Human Services having the most extensive evaluation experience (U.S. Government Accountability Office, 2011). The Department of Education has supported educational research, evaluation, and dissemination since its establishment in 1979. For several years, two

central offices in the Department of Education have been responsible for programme and policy evaluation: (a) The Policy and Program Studies Service (PPSS), in the Office of Planning, Evaluation, and Policy Development (OPEPD), and (b) The Institute of Education Sciences (IES), established in 2002 (replacing the Office of Educational Research and Improvement).

At the Department of Housing and Urban Development program evaluation is primarily centralised in the Office of Policy Development and Research (PD&R), created in 1973. Some evaluation is also conducted by programme offices, such as the Office of Housing, which routinely conducts analyses to update its loan performance models for assessing credit risk and the value of its loan portfolio (U.S. Government Accountability Office, 2011).

Evaluation planning is decentralised in the Department of Health and Human Services. The Department's centrally located Office of the Assistant Secretary for Planning and Evaluation (ASPE) coordinates agency evaluation activities and reports to the Congress, but relies on agencies to evaluate their own programmes. The Centre for Disease Control and Prevention (CDC), as part of the Public Health Service, also supports some evaluation activities. CDC recently created an Office of the Associate Director for Program which will, among other duties, be responsible for supporting performance measurement and evaluation across CDC (U.S. Government Accountability Office, 2011).

Evaluation Capacity-Building in the USA

Professional associations in the USA.

The USA has also taken in the lead in the professionalisation of discipline. For instance, two professional associations for practicing evaluators were founded as early as 1976 (Worthen, 1994). The Evaluation Network (EN) consisted largely of educational evaluators, while most members of the Evaluation Research Society (ERS) served in other professional fields. In 1985, the EN and ERS merged to form the American Evaluation Association (AEA). The AEA's membership has grown from over 3000 members in 2001 to approximately 7000 in 2016.

In 1981, the Joint Committee on Standards for Educational Evaluation (JCSEE) also published comprehensive guidelines to inform the work of evaluators and guide users of evaluation reports (Worthen, 1994).These guidelines were called the Standards for Evaluations of Educational Programs, Projects and Materials. In 1982, the ERS published another set of guidelines for evaluation practice. In 1988, the JCSEE published a second set of guidelines and standards for personnel evaluation.

University-based evaluation training programmes in the USA.

The U.S. Congress funded graduate training programmes in educational research and evaluation as early as 1963 (Worthen, 1994). Examples of USA universities that offered evaluation courses as early as 1963/1968 include: (a) Ball State University, (b) Columbia University, (c) Florida State University, and (d) the University of Chicago (Altschuld, Engle, Cullen, Kim, & Macce, 1994). Prior to the mid-1960s, programme evaluation was taught as a component of research methods courses or as a sub-field of social science disciplines, such as Psychology, Education, and Health. At this stage, programme evaluation had no methodological or theoretical grounding of its own, and as such, each discipline approached evaluation from a different perspective (Altschuld et al., 1994). As the field gained more recognition and theoretical grounding, more distinct training programmes in evaluation were established to train evaluators for the growing demands of evaluation practice. By 1994, there were more university programmes in the USA that offered doctoral degrees in evaluation compared to programmes in Canada and Australia. It should be noted that many existing training programmes are, however, still tied to other social sciences disciplines (Lavelle & Donaldson, 2010). Lavelle and Donaldson (2010) found evidence of 48 university-based evaluation training programmes in the USA in 2008. The study revealed a significant increase in the number of evaluation training programmes in the country, specifically within schools of education. Recent research by LaVelle (2014) found that in 2011–2012, there were over 35 evaluation-specific certificate programmes, 50 evaluation-specific master's degrees, and 40 doctoral programmes in the USA.

USA's Major Contribution to the Field of Evaluation

One of USA's most significant contributions was to provide a conceptual and methodological foundation to the field and distinguish it from other more traditional techniques of accounting and auditing (Derlien, 1990). The need for a conceptual and methodological foundation unique to the discipline was identified by American scholars such as Scriven, Cronbach, Stake and Stufflebeam from early on (Worthen, 1994). By 1970 important seminal writings and publications focusing exclusively on evaluation began to provide conceptual grounding to the discipline. These writings expanded markedly throughout the 1980s and were published in journals such as Evaluation, Evaluation and Program Planning, Evaluation Practice, Educational Evaluation and Policy Analysis, and New Directions for Programme Evaluation.

A number of influential evaluation theories, paradigms and methodologies were developed by American scholars. Shadish, Cook and Leviton (1991) characterised the development of evaluation theories as a series of distinct stages. Stage one theories emphasised the discovery of truth (e.g., the evaluation theories of Michael Scriven and Donald Campbell). Stage two theories focused on evaluation use and social utility (e.g., the theories of Joseph Wholey, Robert Stake, and Carol Weiss). Stage three theory development addressed the integration of Stage one (inquiry) and Stage two (utility) theories (e.g., the theories of Lee Cronbach and Peter Rossi).

In a more recent attempt to systematised evaluation theories, Alkin and Christie (2004) used an evaluation theory tree to depict major theorists and their most significant *Use*, *Methods*, or *Value* contribution. Figure 5 depicts the third version of their evaluation tree (Christie & Alkin, 2008).

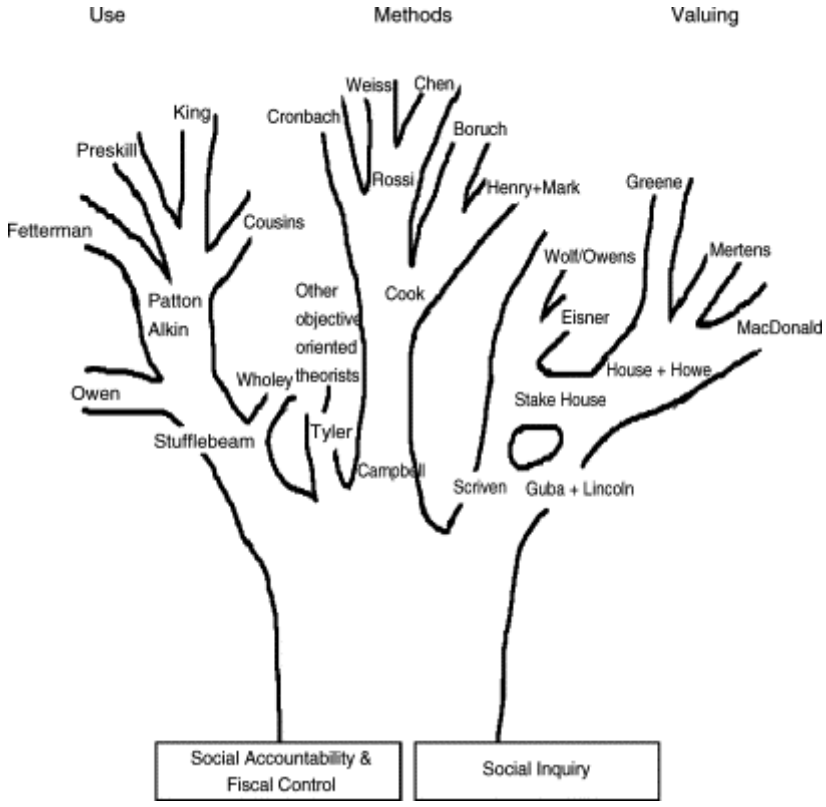


Figure 5. Alkin and Christie's (2006) evaluation theory tree.

The trunk of the evaluation theory tree is built on a dual foundation of accountability and systematic social inquiry. The main branch of the tree is the continuation of the social inquiry trunk and depicts evaluation practice as guided by Methods. The Valuing branch, initially inspired by the work of Micheal Scriven, establishes the vital role of an evaluator in valuing. The third major branch, Use, is inspired by the pioneering work of Daniel Stufflebeam and Joseph Wholey, and depicts evaluation practice as guided by who will use the evaluation information and how. Pioneers in each area (Methods, Use, Valuing) are located at the bottom of each branch and subsequent theorists are placed further alongside the branch.

Many theorists depicted in Figure 5 were at the forefront of numerous seminal developments in the field. For example, Campbell and Tyler advocated the use of a positivist paradigm, governed by experimentation, to guide evaluation designs. Stake (1976), on the other hand, advocated a constructivist paradigm, which favoured qualitative analysis and stakeholders' involvement.

These two perspectives/paradigms were brought together in the form of multi-method approaches, advocated by theorists such as Greene. The highly contested black-box evaluation approach was also first addressed by USA theorists such as Chen and Rossi (1989). Other more recent scholarly contributions by USA theorists include a discussion of the links between evaluation theory and practice and how to bridge the gap between theory and practice (Fitzpatrick, Christie, & Mark, 2009).

Summary

This sub-section first provided an overview of how evaluation progressed in the USA government at both executive and legislative levels and offers a tentative assessment of the current status of evaluation in the US. The brevity of this description precludes in-depth and nuanced discussion of the different ways in which the evaluation function is carried out among the different executive branch departments or the committees and subcommittees of the U.S. Congress. As highlighted in this sub-section, evaluation is rarely singled out as the sole function of any governmental agency in either branch (Melkers & Roessner, 1997). Instead, the practice of programme evaluation is scattered within both the legislative and executive branches. On the whole, evaluation is firmly institutionalised within the democracy and has reached a high level of methodological sophistication. The USA has one of the most advanced evaluation systems. Furthermore, the USA played a major role in influencing the skills and willingness of other countries to introduce programme evaluation and further develop the discipline (Rist, 1990).

The USA is considered to be a forerunner in the discipline of programme evaluation. For comparison purposes, it is worth noting that by the time evaluation emerged in Brazil and South Africa, scholars in the USA had already been engaging in debates around the legitimacy of the field, designed different training options and formulated a number of evaluation theories and approaches (Mouton, 2010).

THE UNITED KINGDOM (UK)

The UK's Key Historical and Political Markers

The UK is a constitutional monarchy and parliamentary democracy. Constitutional features include: no written constitution; an institutionalised adversarial parliamentary system; and a complex central-local government linkage, in which some services (e.g., health and education) are administered centrally but delivered locally (The Commonwealth, 2016).

With a GDP of $ 2.849 trillion, the UK is the third largest economy in Europe after Germany and France (The World Bank, 2016). After emerging from the economic recession in 1992, the country's growth rate consistently outpaced most of Western Europe. In 2008 however, the global financial crisis ended the 15 year period of continuous growth and stable inflation, pushing the economy into an unprecedented recession. This recession, triggered by the banking crisis brought "the longest boom' on record to an abrupt halt" (Martin, Gardiner, Tyler, 2014, p. 14). Recovery from this last recession has been a long process. It took six years for output to return to its pre-recession peak in 2008.

The UK's Current Socio-Economic Standing

The UK has a strong presence in international affairs, and considerable political and cultural influence on many countries (The World Bank, 2016). As one of five permanent members of the UN Security Council and a founding member of NATO and the Commonwealth, the UK adopts a global approach to foreign policy.

The UK performs relatively well in terms of key socio-economic indicators. The country's employment rate and average earnings is higher than the OECD average (OECD, 2016d). The unemployment rate is about 5%, and the unemployment benefit claims have declined by more than 50% since 2013 with the introduction of the "Help to Work" programme (OECD, 2016e).

Educational attainment among the adult working-age population and expenditure per student by primary, secondary and tertiary institutions is higher than the OECD average (OECD, 2015c). Personal safety is also relatively high in the UK, with the country having one the of lowest homicides rate amongst OECD countries. High-quality health care is also a key priority in the UK (even if Health spending per person is slightly below the OECD average) and access to health care is generally good (OECD, 2015b).

The UK's Current Socio-Economic Challenges

The country's economic growth is projected to be only 1.75 % in 2016. Membership of the European Union (EU) had contributed to the economic prosperity of the UK, but with UK exiting from the EU (Brexit), projections are bleak. Kierzenkowski, Pain, Rusticelli and Zwart (2016, p.16) argue that "Brexit would be akin to a tax on GDP, imposing a persistent and rising cost on the economy that would not be incurred if the UK remained in the EU". In the short-term term, the economy is expected to suffer from tighter financial conditions and weaker confidence. After formal exit from the EU, the country will experience higher trade barriers and restrictions on labour mobility. The UK's current trade deficit of 7% is also the highest on record, thus compounding the projected economic downturn.

Other socio-economic challenges include: (a) relatively poor child health indicators for an OECD country, with, for example, 3.1% of British children being obese; (b) stark regional income inequalities, with the average household adjusted disposable income almost 60% higher in Greater London than in Northern Ireland; (c) uneven quality of healthcare and a financially strained National Health System (OECD, 2016d).

The Emergence and Development of Programme Evaluation in the UK

Evaluation emerged in the European community, at the beginning of the 1980s, in the context of the New Public Management (NPM) approach (Stame, 2003). Many European states introduced systems of NPM and programmes for social rebalancing during this period. These new initiatives were pivotal for the development of

programme evaluation in Europe as the demand for evaluation activities initially arose “in a spirit of obedience to central conjunctions” (Toulemonde, 1995, p.85). In the 1990s, evaluation became widespread across a number of European countries. In almost every European country, aspects of NPM were introduced as part of public sector reforms (Stame, 2003). The demand for evaluation spread across Europe as a parallel process.

The history of programme evaluation in the UK can be traced back to the early 1960s. Evaluation, more specifically policy evaluation, progressively developed within the British central government in line with the government’s agenda to control and prioritise public expenditure allocations, as well as raising the profile of public management by advocating systematic and rational decision-making (Gray & Jenkins, 1982). The Public Expenditure Survey System (PES) introduced in the early 1960s was specifically designed to plan and coordinate public expenditure allocations on a more rational basis. The PES however lacked the analytical capability to monitor and evaluate the impact of policies. A series of reforms in the early 1970s, implemented during the Conservative regime led by Heath, sought to address this shortcoming. These reforms included the introduction of Programme Analysis and Review (PAR) and the Central Policy Review Staff (CPRS) cabinet (Derlien, 1990; Gray & Jenkins, 1982). PAR was an approach to policy analysis, introduced as part of a systematic attempt, both in the UK and elsewhere, to develop rational management within the government structure and to transfer ideas developed in private sector management into public sector organizations (Gray & Jenkins, 2002). The PAR was introduced to fulfil three broad aims: (a) to enhance strategic management within the government, (b) to increase ministerial oversight over departmental activities, and (c) to provide an evaluative mechanism for the annual PES. These three aims reflect the conservative government’s strategic emphasis on departmental performance and delivery and public expenditure control. Two distinctive features of the PAR are worth noting here:

- While the PAR was coordinated at the centre, it was administered by individual departments.

- The PAR was concerned with the appraisal of policies rather than of programmes (Derlien, 1990).

The CPRS, on the other hand, was a cabinet think tank established to assist ministers to align policy and programme decisions with long-term goals and assess the cumulative impact of both policies and programmes. Both the PAR process and CPRS's work encountered a number of technical, organizational and political barriers, that were difficult, if not impossible, to overcome. While the initial reviews conducted between 1971 and 1973 were driven with enthusiasm by government officials, a number of factors undermined the exercise from 1973 onwards (Gray & Jenkins, 2002). In particular, there was a shift in the government's strategic priorities in the mid-1970s, with more political commitment devoted to careful management of public expenditure than to policy analysis and evaluation given the dire fiscal situation and the financial crisis that prevailed during this period (Derlien, 1990). The early evaluation impulse gradually faded as a Labour government came into office in 1974, first under the leadership of Wilson and then Callaghan in 1976. With pay-offs seen as minimal, the PAR was formally dismantled in 1979 (Derlien, 1990).

The impulse for evaluation was reactivated when the Conservative government was elected into office in 1979, under the leadership of Thatcher (Derlien, 1990). While the emphasis remained on resource management and tight fiscal control, Thatcher's managerialist approach and support for the Value for Money movement led to the resurgence of policy evaluation. In addition, the need for state auditing and enhanced regulation led to the establishment of a number of national agencies, such as the National Audit Office (NAO), and the Audit Commission (AC). Both agencies were established in 1983. The NAO was established to serve the Public Accounts Committee (PAC) (Derlien, 1990). The agency conducted Value for Money (VFM) studies, with a particular emphasis on issues of efficiency and effectiveness. Although these studies mimicked evaluation-like activities to some extent, by assessing programme objectives, the work of NAO predominantly remained financial. The agency was also tasked with auditing the accounts of central government departments and their agencies, as well as undertaking certification audits. The primary function of the AC, on the other hand, was to regulate the activities of local authorities and ensure that they implement Value for Money

principles in the management of public funds (Kelly, 2003). Unlike the NAO, the AC's activities often extended beyond the financial issues to consider implementation challenges and failure. Both the NAO and the AC and are still in operation.

By mid-1980, a number of developments pointed to the revival of performance monitoring and evaluation in the UK government. For example, a small team of Treasury and Cabinet Office officials formed the Joint Management Unit (JMU) in 1985 (Derlien, 1990). The JMU was tasked with the development of an evaluation system for use in over 20 departments and across a variety of programmes. The JMU's activities were critical to the development of departmental capacity to plan and conduct policy and programme evaluation on a regular basis. In the late 1980s, the central government also began to strengthen performance reporting requirements of various public entities. For instance, all newly established civil service executive agencies were required to report on key performance indicators (KPIs) on an annual basis, from 1988 onwards (Talbot, 2010).

Despite those developments, evaluation activity in the UK was fragmented and seen as a tool for expenditure reduction in the late 1980s. There was no established community devoted to policy evaluation and formalised procedures for initiating, conducting and utilising evaluations in the policy process were limited at that time. This situation persisted throughout the 1990s. As a result of the tight fiscal conditions that continued under the Conservative leadership of Major, the emphasis remained on measurements of efficiency and throughput, and the increasingly active participation of regulators. There were no fewer than 134 separate bodies regulating the UK public sector at national government level in 1995 (Hood, James, Jones, Scott, & Travers, 1998). All in all, there was however a withdrawal from traditional programme evaluation until the New Labour government was elected in May 1997 (Grey & Jenkins, 2002).

One of the first developments that indicated a movement towards a stronger evaluation focus was the introduction of Spending Reviews (SRs) and Comprehensive Spending Reviews (CSRs), by the then-Chancellor Brown in 1997 (Mullard, 2001; Talbot, 2010). The CSRs are thorough strategic reviews of spending

priorities, while SRs culminate into incremental and medium-term changes to existing priorities. Each CSR and SR set Departmental Expenditure Limits (DELs) for ministries over a three year period (Talbot, 2010). Ministries are then required to set clear and quantifiable objectives and targets as part of their Public Service Agreements (PSAs) (Gray & Jenkins, 2002). PSAs were agreements negotiated between Treasury and spending ministries that stipulated the allocation of resources in exchange for delivery. The first set of PSAs was published in 1998 and each subsequent CSR/SR was driven by a revised set of PSAs. The first CSR was published in 1998, followed by three spending reviews in 2000, 2002, and 2004, and another CSR in 2007 (Talbot, 2010).

Other initiatives launched post 1997 under Blair's New Labour administration further strengthened performance monitoring in the UK. These initiatives, collectively known as the local government modernisation agenda (LGMA), represent a far-reaching attempt to enhance the performance of local authorities in the UK and increase public confidence in government institutions (Hood, James, & Scott, 2000; Martin, 2002). Some of the key elements of this agenda draw heavily on New Public Management principles and build directly upon previous reforms that encouraged local authorities to rely on external monitoring. The LGMA is supported by three tranches of legislation. The first one is the Local Government Act passed in 1999 (Martin, 2002). This Act introduced the Best Value regime, under which local authorities were mandated to:

- develop a corporate strategy that outlines their objectives, how these will be pursued and the criteria against which success will be measured;
- undertake performance reviews to examine the purpose of every function and assess the cost effectiveness of alternative approaches to service delivery; and
- publish annual performance plans specifying strategies and targets for improvement and criteria for monitoring progress.

The Local Government Act 1999 also consolidated the formal duties and powers of the AC. The 1999 Act requires the AC to prepare and review the Code of Practice for

the audit of local authorities and assess their Best Value Performance Plans. The Act provides the AC with the necessary powers to assess whether local authorities are complying with and fulfilling their obligations (Kelly, 2003). A second Local Government Act was introduced in 2000, requiring local authorities to develop community strategies, new political management structures and a revised code of conduct.

A third phase of local government reform, involving a differentiated framework for the regulation of local councils' activities, was introduced following the publication of the 2001 Local Government White Paper (Martin, 2002). Local Public Service Agreements (LPSAs) were established between individual local authorities and the central government. The LPSAs scheme is an extension of the PSA system promulgated since 1998 (Talbot, 2010). Under the LPSAs scheme, additional government funding is contingent upon, and relative to the degree to which local authorities achieve improvement targets, thus indicating a strengthened focus on evaluation-led allocation of resources (Martin, 2002).

PSAs have evolved considerably over the five iterations that have been published so far. From the second iteration, PSAs began to focus more on measures of outcomes than on processes. The biggest qualitative change to the PSA system was implemented in 2007 with the second Comprehensive Spending Review. From 2007, all PSAs became cross-cutting in nature, with targets being shared across two or more Ministries. In 2007, each PSA was underpinned by a Delivery Agreement, which was shared by relevant Ministries. Departmental Strategic Objectives (DSOs) were also introduced to supplement the PSAs. The DSOs, in principle, served as quasi-contractual agreements between the Treasury and the Ministries (Talbot, 2010).

A number of recent developments within the UK government indicate a strengthened focus on evaluation and increased commitment to independent evaluation. These include the establishment of the Independent Commission for Aid Impact (ICAI) in 2011 to evaluate the impact and value for money of the UK's aid/ Official Development Assistance (ODA) programmes. The ICAI is an independent body that reports directly to the International Development Select Committee in Parliament

(Independent Commission for Aid Impact, 2011). The Department for International Development (DFID), which manages around 86% of all UK's ODA, also embeds evaluation as part of the design of its aid programmes. The DFID has a results framework that monitors progress against key development outcomes. The framework also includes key performance indicators that monitor DFID's operational effectiveness (Department for International Development, 2015).

Key Players in Policy Evaluation in the UK

There are a number of sources of advice within the UK government about evaluation policy and practice. The Department of Work and Pensions (DWP), in conjunction with the National Centre for Social Research, for example, published the Research Methods for Policy Evaluation brief in 2001 to highlight the main evaluation methods used within the DWP for evaluating active labour market programmes and policies (Talbot, 2010).

The Treasury has also, for many years, played a key role in strengthening the practice of evaluation within the central government, by providing guidance to departments and executive agencies on how proposals should be appraised before resources are allocated, and how programmes, projects and policies should be evaluated upon completion. One significant output of the Treasury is the publication of the Green Book, which serves as reference guide for departments and executive agencies throughout the appraisal and evaluation process. The Green Book was last updated in July 2011 (HM Treasury, 2011). In addition, the Treasury, in conjunction with the Cabinet Office, the NAO, the AC, and the Office of National Statistics, published an agreed set of definitions and concepts relating performance monitoring in order to minimise terminological confusion within the government (Talbot, 2010). The NAO and AC have been particularly active in developing capacity within the government to monitor and interpret performance data.

The Government Social Research Service (GSRS) has also played an important role in the development of evidence-based policies. The GSRS is a professional network of social researchers and evaluators across government. The GSRS's main role in relation to M&E is the publication of the Magenta Book, which is a collection of

guidance notes on policy evaluation and analysis for evaluation practitioners (Talbot, 2010).

Evaluation Capacity-Building in the UK

The UK Evaluation Society (UKES).

The UKES was founded in 1994 to promote and improve the theory, practice, and utilisation of evaluation. UKES has networks in five different regions, namely Wales, the North West, the North East, the Midlands, London and the South West. The UKES currently has a diverse membership comprising evaluation professionals and practitioners from national and local governments, independent consultancies and the voluntary sector. The society is affiliated to the European Evaluation Society and the International Organization for Cooperation on Evaluation (IOCE). The society hosts an annual two-day conference, pre-conference practitioner workshops, an annual national training event, and a number of regional seminars. The society also publishes a newsletter, The Evaluator, thrice a year (UKES, 2016).

University-based evaluation training programmes in the UK.

There is no systematic online listing of university-based monitoring and programme evaluation training options in the UK. Nine postgraduate courses in evaluation were identified by Davies (2008). This list was last updated in 2004. The profile of one university (London Metropolitan University), offering an MSc in Social Research and Evaluation features in a listing compiled by University of Bern's Centre for University Continuing Education in 2012. The profiles of sixteen other European universities (based in Belgium, Switzerland, Germany, Great Britain, Italy, France, Spain, Sweden, Netherlands, Greece, Romania, and Denmark) also feature in this listing. A few examples of universities offering monitoring and evaluation related courses in the UK include: University of Southampton (MSc Environmental Monitoring and Assessment), University of Manchester (MSc Management and Implementation of Development Projects), and University of Oxford (MSc in Evidence-Based Social Intervention and Policy Evaluation).

Other evaluation capacity-building organizations in the UK.

In addition to attending seminars and workshops organised the UKES and its affiliates, M&E related knowledge and skills can be acquired through a number of M&E training providers in the UK. These include Charities Evaluation Services (CES), The International NGO Training and Research Centre (INTRAC) and IMA International (Davies, 2008). Both CES and IMA offer a range of training courses, including a foundation courses in monitoring and evaluation, data analysis and reporting of evaluation findings. INTRAC provides training, consultancy and research services to organizations involved in international development and relief, including NGOs. INTRAC's training programme concentrates on issues around strengthening civil society, organizational capacity building and programme development (Davies, 2008).

UK's Major Contribution to the Field of Evaluation

The realistic evaluation approach is arguably UK's most significant contribution to evaluation theory. This theoretical approach emerged towards the late 1990s and was first presented in Pawson and Tilley's *Realistic Evaluation* (1997). The approach provides a distinctive and multi-faceted account of the nature of programmes and how they work (Pawson & Tilley, 2004). The main aim of realist evaluations is to produce a tested theory about what works for whom in what circumstances, and in what respects. Realist evaluation is applicable in principle to all forms of programme evaluation, and to all areas of social and public policy (Pawson & Tilley, 2004).

Summary

Compared to countries such as Switzerland, Germany and France, the UK has a longer evaluation tradition and belongs to the group of European countries that launched the evaluation function early on (Widmer, 2004). It should however be noted that government policies in the UK have focused mainly on performance monitoring, rather than on evaluation. Despite definite increases in evaluative activity after the New Labour government power in 1997, performance monitoring remains the predominant approach within the government (Talbot, 2010). It is however clear

that the UK has successfully developed a comprehensive performance monitoring system that has become increasingly sophisticated and outcomes focused over time. Virtually all UK public entities have either legal or administrative mandates to produce publicly available performance that, in principle, inform resource allocation and decision-making.

Rationale for Selecting Evaluators from Brazil and South Africa as Units of Analysis

Broadly speaking, both Brazil and South Africa are developing countries, with similar historical, social and economic trajectories (Nayyar, 2008). In 1998, Brazil emerged from a twenty year military dictatorship and embarked on a democratisation process. South Africa, on the other hand, reached full democratic state in 1994, after forty years of apartheid rule. Brazil is the dominant country of South America, with a population of 207.8 million people and a GDP of US $ 1.775 trillion in 2015. South Africa is analogously the second wealthiest country in Africa, with a GDP of US $ 312.8 billion in 2015. Both countries currently have one of the most robust and extensive welfare states in the developing world, and are in a crucial stage of transition towards democratic practices, transparency and reflection. In addition, both countries are faced with similar socio-economic and developmental challenges, including unequal income distributions, low economic growth rates, and relatively poor educational outcomes (Maia, Mondi, & Roberts, 2005). At the same time, both countries are attempting to chart a progressive economic path, while being important role players in the global economy. Both countries form the core of integration programmes in their respective regions–Brazil for Mercosul (the Common Market of the South) and South Africa for SADC (Roelofse-Campbell, 2006). Both countries form part of the BRICS nations, along with other emerging powers - Russia, India, and China.

Programme evaluation also emerged around the same time in both Brazil and South Africa and as an inherent part of the re-democratisation process in the 1990s. Both countries have made significant progress in evaluation policy and practice. The advances produced in the area of M&E over the past few years, are unquestionable in both countries. Brazil is one of the few South American countries that have a

robust M&E system. Similarly, South Africa is one of the leading countries advancing the M&E agenda in Africa. Both countries however face similar challenges in terms of evaluation expertise and training. The existing evaluation knowledge is not spread and systematised to the desired level in Brazil and the training of evaluation professionals still remains a challenge (Firme et al., 2009). Evaluation culture is also not yet fully developed. This situation mimics the current evaluation landscape in South Africa.

Given that Brazil and South Africa share key similarities in terms of their socio-political context, and current state of programme evaluation, one can reasonably expect that patterns of practice of South African evaluators and Brazilian evaluators to be comparable. This study will test this assumption and contrast the pattern of evaluation practice in these two developing countries with that of two developed countries that have a more advanced M&E system and structure: the US and the UK.

Rationale for Selecting Evaluators from USA and UK as Units of Analysis

Both the USA and the UK have a constitution-based government and are comparable in terms of developmental stage, with the UK being the second largest economy in Europe, and the USA being the world's leading economic power. In addition, the USA and the UK are bound by a common language, and have well-established trade and economic relations. On the diplomatic front, both countries are among the founders of the United Nations, NATO, the World Trade Organization, G-8, and a host of other international bodies.

A number of parallels can also be drawn between the UK and USA in terms of how programme evaluation evolved in these two different countries. In both the UK and the USA, the central government has been the primary driver of the discipline. In both countries, auditing institutions, the GAO in the USA and the AC and NAO in the UK, were initially tasked with the evaluation function. In both countries, the agenda of various political administrations influenced the evolution of and the importance attached to the field. For example, in the USA programme evaluation gained more grounding under Democratic administrations than under Republican administrations.

Similarly, in the UK, programme evaluation stagnated under the Conservative leadership of Thatcher but experienced an upsurge in 1997 with Blair's LGMA. Furthermore, the prevailing fiscal situation in both countries dictated the scope and purpose of evaluations. For example, in the mid-1970s, evaluation activities were scarce in both countries due to tight fiscal conditions. When conducted, evaluations were used to inform and justify resource allocation.

Even though the USA and the UK share key similarities in terms of how programme evaluation evolved, it should be noted that the discipline developed at a slower pace in the UK compared to the USA. For example, despite the early institutionalisation of the discipline in the UK, guidelines and standards for good practice were only developed in 2003 by the UKES (Widmer, 2004). A long tradition of exchange, however, exists between British evaluators and American evaluators (Wider, 2004). In addition, both American and British evaluators practice in a developed economy. One can therefore reasonably expect the pattern of practice of British and USA evaluators to be comparable. This study will test this assumption.

Conclusion

The key and defining criteria for selecting Brazil, South Africa, UK, and USA to recruit the sample of interest is that all four countries can be neatly clustered as a developed or developing country, with or without a mature evaluation culture. Brazil and South Africa entered the evaluation era in the 1990s and are considered newcomers in the evaluation community. In both countries, the evaluation culture was externally driven. On the other hand, the evaluation culture in the USA was internally driven. The UK is among the four countries (Canada, Sweden, Germany) that adopted an evaluation culture in the 1960s/1970s, and is thus regarded as an early adopter. The UK evaluation culture was strongly influenced by the American discourse on evaluation. Both USA and UK are considered as countries with mature evaluation cultures, and strong evaluation capacity.

The distinction between these two clusters of countries, the early adopters and the newcomers, gives rise to questions about differences in evaluation practice, more specifically:

How do evaluators practising in countries that adopted evaluation in the 1990s differ in terms of patterns of practice compared to those practising in countries where an evaluation culture has existed for decades?

The particular pattern of practice investigated in this study relates to decisions about programme evaluability. More specifically, one of the aims of this study is to isolate the criteria that evaluators from Brazil, South Africa, USA and UK prioritise when assessing the evaluability of a programme.

CHAPTER FOUR

METHOD

The simulation study aimed at investigating whether evaluators with different characteristics share a common and consistent perspective towards evaluability. Evaluators from four different countries were presented with three fictitious evaluation scenarios and a Q Sort task. This simulation design allowed for the direct comparison of evaluator reactions to different evaluation scenarios and pre-determined evaluability criteria, and provided insight into whether or not evaluators' characteristics predicted their evaluability decisions.

The following three inter-related research questions guided the investigation:

1. Do evaluators share a common perspective towards evaluability? If not, what perspectives can be empirically identified and what evaluator types are most associated with these perspectives?
2. Are evaluators' prioritisation of evaluability criteria consistent across different study tasks (three different evaluation scenarios, and one a-contextual sorting task)?
3. Do selected evaluator characteristics (practice context and experience) predict their evaluability assessments, likelihood to evaluate, and prioritisation of evaluability criteria?

This chapter reports on the method used to answer the research questions and presents the rationale underlying each of my methodological choices. It covers four main sections: study design, measures, participants and the procedure used to collect and analyse the data for both the pilot study and the main study.

Design

A descriptive design was used for this study. According to Babbie and Mouton (2001) "description is the precise measurement and reporting of the characteristics of some population or phenomenon under study" (pp. 105). The phenomenon under

study in this case was evaluability and the population consisted of evaluators in four different countries. The use of a descriptive design was deemed appropriate as I was primarily concerned with the collection and description of cross-sectional data relating to participants' characteristics and patterns of practice as opposed to answering questions about how or why these patterns occurred.

Measures

Three distinct measures were used to capture the variables of interest: a Q Sort task, three evaluability scenarios, and close-ended items relating to evaluator characteristics. These measures were embedded in an online survey. A web-based data collection method was chosen for a number of pragmatic reasons, including:

- Ease of administration.
- Accuracy of data entry.
- I could minimise the extent of missing data by programming the survey in such a way that all items in a given section had to be answered before the next section could be accessed.
- I could ensure that participants did not take the survey multiple times. A unique survey link was automatically created for each participant.
- I could embed randomisation processes in the survey, where appropriate.

The design of the online survey proceeded as follows:

Coversheet.

A coversheet containing information about the nature of the study and details relating to the researcher, her supervisor and their academic affiliation was designed to ensure that potential respondents were able to make an informed decision regarding participation. A statement on ethics clearance (see Appendix B) and an undertaking to keep responses anonymous were included. The survey link and a request to forward the study invitation to other eligible evaluation practitioners were also included. The eligibility criteria and incentive for participation were clearly articulated.

Q sort task.

A Q Sort task (see Appendix C) was used to answer the first two research questions. The primary aim was to identify systematically dominant patterns that may arise among evaluators in terms of how they prioritise evaluability criteria. The secondary aim was to determine the distinct profile of participants exhibiting a particular pattern.

The Q Sort task was designed following the principles of the Q Sort method (Stephenson, 1935). The Q Sort method is a procedure that facilitates the systematic study of participant subjectivity (Cross, 2005). An important premise underlying this method is that, just like any other behaviour, subjectivity can be systematically analysed when expressed in operational terms.

In a Q study, participants are presented with a set of randomly ordered statements relating to a specific topic and are asked to sort these statements into a subjectively meaningful pattern based on their individual preference or judgement. Depending on the condition of instruction, participants might be required to sort each statement into a normalised grid (Ramlo, 2005). This procedure is called Q sorting. The individual rankings are then subject to a factor analysis (van Exel & de Graaf, 2005). A defining feature of the Q method is that statements relating to the same domain are not analysed individually but in the context of other equally relevant statements.

The Q factor analysis is an inversion of conventional factor analysis (R factor analysis) in the sense that Q correlates personal profiles instead of items, thus providing information about similarities and differences in viewpoints on a specific topic. In a Q factor analysis, participants who complete the Q Sort task are equivalent to the variables in a conventional factor analysis (Cross, 2005). The Q factor analysis allows researchers to extract different segments/clusters of subjectivity (factors) among participants, identify dominant viewpoints or preferences, and isolate the defining characteristics of participants who subscribe to different factors (Eghbalighazijahani, Hine, & Kashyap, 2013; van Exel & de Graaf, 2005). This type of classification is often referred to as typology development (Ramlo & Newman, 2011).

Q studies have often been criticised as small sample investigations of subjectivity based on sorting of items of unknown reliability (Thomas & Bass, 1992). Such scepticism is unwarranted for two main reasons. First, the Q method attempts to isolate subjective structures in the data and the extent to which these are similar or dissimilar, rather than calculating the percentage of the sample or population that adheres to them. The Q method does not attempt to estimate population statistics (Stainton Rogers, 1995). It is used to identify a typology and not test the typology's proportional distribution within the larger population (Eghbalighazijahani et al., 2013). The issue of generalisability is therefore not relevant here.

The Q method aims to explore the different accounts that people construct. According to Stainton Rogers (1995), the focus of this method is not on the constructor (i.e., the participants) but on the constructions themselves. Since factors represent qualitative categories of thought, additional participants would have minimal to no impact on the factor scores (Brown, 1993). The issue of large sample sizes is therefore relatively unimportant in Q studies. According to Brown (1980) only a limited number of distinct viewpoints exist on any given topic and a well-structured Q sample, containing a wide range of existing opinions on the topic, should reveal these perspectives.

Second, the most relevant type of reliability that applies to this method is replicability. According to Brown (1980), a Q sort can be replicated with 85% consistency up to a year later. It should however be noted that the Q method does not necessarily yield the same results when repeated on the same participants on two separate occasions. Stainton Rogers (1995) argues that this is not problematic as there should not be an expectation that an individual will express the same views on different occasions as views evolve with time. What a researcher should be concerned about is whether the same condition of instruction will lead to factors that are schematically reliable and represent similar viewpoints across similarly structured yet different Q samples (Van Exel & de Graaf, 2005).

I chose to use the Q method in this study as it is particularly well suited to studying phenomena "in which there are numerous ideals present in a reality where only a

limited number of ends or means can be realistically pursued" (Thompson, 1998, p.1). This is particularly evident in an evaluation context, where practical realities constrain evaluators to prioritise a set of evaluability criteria at the expense of others. Using this empirically robust method allowed me to capture evaluators' perspectives without having to conduct lengthy interviews.

Application of the Q method in the field of programme evaluation is sparse. A notable exception is a study conducted by Thompson and Miller (1983), which investigated administrators' and evaluators' perceptions of programme evaluation. Authors, such as Ramlo and Newman (2011), have called for more researchers in the field to use this method to derive stable evaluator and stakeholder profiles.

The design and implementation of the Q Sort task for this study involved the following steps: (1) definition of the concourse, (2) development of the Q set, (3) specification of the Q Sort protocol and response format, and (4) selection of the P set. Each of these steps are discussed below.

Definition of the concourse.

In the Q Sort method, concourse refers to the flow of communicability surrounding a particular topic, that is, the existing opinions and perspectives around the topic. The level of discourse around the topic dictates the sophistication of the concourse (Brown, 1991). The concourse in essence comprises the raw material for the Q method. It is from this concourse that the Q statements are derived for subsequent administration in the Q sort exercise.

As starting point, I conducted a literature review of the existing opinions and perspectives of evaluation theorists and practitioners on the concept of evaluability. The discourse on evaluability is presented in Chapter 2 and summarised in Appendix A.

Development of the Q set.

The Q set consists of a subset of statements drawn from the concourse. These statements represent matters of opinion as opposed to facts (Brown, 1991). According to Brown (1980), it is critical to derive a Q set that is representative of the wide range of existing perspectives around the topic. Propositions embedded in the concourse need to be carefully sampled as participants cannot construct a meaningful story if the appropriate statements have been not been included in the Q set (Stainton Rogers, 1995).

I used the concourse presented in Chapter 2 to derive a model of evaluability. This model is presented in Figure 4. The context block was first disaggregated into 19 distinct evaluability criteria and further refined and categorised under either programme features, stakeholder characteristics, or logistical requirements. Each evaluability criterion was formulated as a statement. The development of the Q set is described in full in Chapter 2. The Q set statements used in this study are also presented in Chapter 2 (see Table 5).

Specification of Q sort protocol and Q response format.

A researcher needs to articulate the condition of instruction to facilitate the sorting process and decide whether to use a forced-choice or a free-sort condition of instruction (Du Plessis, 2005; Watts & Stenner, 2005). Most Q studies require participants to place a predetermined number of Q statements into a predetermined number of categories. This protocol uses a forced-choice response format in which placement of one Q statement into a given category inherently constrains the possible placements of subsequent statements. This response format yields a normal or quasi-normal distribution of scores for each participant (Thompson, 1998). Figure 6 presents the forced distribution of 24 Q statements on a normalised grid.

Disagree			Neutral / no particular opinion			Agree
-3	-2	-1	0	1	2	3

Figure 6. Forced distribution of 24 Q statements.

A researcher can also adopt a two-stage measurement protocol, whereby participants are first required to follow a Q Sort protocol that does not impose any restrictions in terms of the number of Q statements that could be placed in a given category (free-sort condition of instruction). Participants may place any number of Q statements under the distribution markers until all Q Sort items have been sorted (Du Plessis, 2005; Watts & Stenner, 2005).They are then required to rank-order the Q statements within each of the Q Sort categories. This particular protocol yields more variance in responses, and hence facilitates the extraction of more stable factors (Thompson, 1998).

I used a forced-choice condition of instruction in the pilot study and adopted a two-stage measurement protocol and a free-sort condition of instruction in the main study. The rationale for using this approach will be explained later in this chapter.

Irrespective of whether a forced-choice or a free-sort condition of instruction is specified, the appropriate response format needs to be selected. The response format represents the labels assigned to the ranking dimensions along which participants sort the Q statements. For example, if the subjective area of interest is

agreement, participants would be required to sort the Q statements along a continuum of *most disagree* and *most agree* (Du Plessis, 2005).

In this study, the ranking dimensions ranged from not all important to essential, as I was interested in how the Q statements were prioritised.

Selection of the P set.

The P set is a structured sample of respondents who are theoretically relevant to the problem under consideration (van Exel & de Graaf, 2005). These respondents must be purposively selected with the expectation that they possess clear and varied viewpoints on the topic under investigation (du Plessis, 2005). The number of respondents is therefore of less importance than who they are. There is in fact no clear cut rule as to how many participants should be included in the P set. Inconsistent suggestions have been made in this regard by proponents of the Q method (Dziopa & Ahern, 2011).

A study conducted by Eghbalighazijahani et al. (2013) discredited previous studies emphasising that the P set must be smaller than the Q set (i.e., the number of participants must be less than the number of Q statements). These researchers argued that: (a) as long as an acceptable amount of variance can be explained by a reasonable number of factors, increasing the number of participants can be useful; (b) the Q method is equally suitable for a small number and a large number of participants; and (c) it is difficult to specify the number of participants required for Q studies as this calculation depends on variables such as the number of Q statements, expected or desired number of factors, total explained variance, and the level of consensus among participants on the topic, which in turn influences the number of factors extracted.

The P set in this study consisted of evaluation practitioners from four different countries (Brazil, South Africa, UK and USA). Most of these evaluation practitioners were affiliated to one of the following professional associations: BMEN (Brazil), SAMEA (SA), UKES (UK), or AEA (USA). The eligibility criteria and characteristics of the P set are described in full under *Participants*. This particular cohort of

participants was purposively selected as they are expected to possess unambiguous and varied views on evaluability.

Evaluability scenarios

Evaluability scenarios were used to answer the third research question. The aim was to assess whether evaluators would use the same evaluability criteria that they prioritised in the Q Sort task to guide this decision making process, when confronted with scenarios that mimic real evaluation situations. In line with authors such as Tourmen (2009) and Azzam (2011), this study used a simulation design to examine evaluator practice in a simulated context. According to Tourmen (2009) the logic of evaluative practice is embodied in an activity flow and often remains implicit. In order to characterise, describe and analyse this logic it is important "to make the tacit explicit" (pp. 8). One way to do this is by simulating a situation that embodies key elements of real evaluation situations, and asking participants to respond to a series of questions aimed at eliciting their decision making template, thus increasing both experimental and mundane realism (Crano & Brewer, 2002). A simulation design was also utilised in this study as it allowed me to keep certain contextual factors consistent.

Three evaluation scenarios were created (see Appendix D). Each scenario included a fictitious description of a programme and a specific set of evaluability conditions. The programme area, educational support for post-high school students, was standardised across all three scenarios in order not to contaminate the scenarios with programme areas in which evaluators may not have expertise. The level of detail in the scenarios was not comprehensive but incorporated the key elements that are typically reflected in Requests for Proposals (RFP).

Ideally, the three iterations of the programme scenario should have manipulated all 19 evaluability criteria derived in Chapter 2. However, as this was practically impossible, a decision was taken to collapse these criteria into three broad categories, namely programme structural features, stakeholder characteristics, and logistical requirements. These three broad categories were manipulated so that three types of scenarios emerged (see Table 12).

Table 12

Evaluability Scenarios

Scenario 1	Scenario 2	Scenario 3
Robust programme features	Weak programme features	Weak programme features
Unfavourable stakeholder characteristics	Favourable stakeholder characteristics	Unfavourable stakeholder characteristics
Unfavourable logistical conditions	Unfavourable logistical conditions	Favourable logistical conditions

Each scenario embodied one favourable and two unfavourable evaluability categories in order to mimic the nature of real evaluation situations. The design of the scenarios was based on the premise that: (a) no evaluation situation is perfect (i.e., each evaluation situation has, in principle, a unique set of inherent challenges); (b) no evaluation situation is completely imperfect (i.e., no evaluation situation is so challenging that some form of evaluation cannot be attempted); and (c) decision making in the context of challenging situations is more nuanced, thus making it easier to isolate and analyse differences in practice.

A within-subject design was used for the scenario task. The independent variable in this case was the nature of each scenario. The dependent variables included participants' assessment of the level of evaluability and the likelihood of evaluating the programme depicted in each scenario. The treatment consisted of exposure to scenarios 1, 2 and 3. The design of the scenario task is presented below:

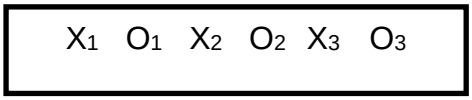


Figure 7. Design of scenario task.

As illustrated in Figure 7, participants were repeatedly measured on the dependent variables (O_1, O_2, O_3), after each treatment exposure (X_1, X_2, X_3). Each participant received the treatment in a randomised order.

Pilot Study

The purpose of the pilot study was to: (a) assess the level of realism of the evaluation context and scenario descriptions, (b) assess whether or not each scenario embodied the intended manipulations, and (c) test the functionality of the web-based survey.

Four evaluators with both an academic interest and actual experience in programme evaluation were purposively selected for the pilot study. The aim was to generate targeted qualitative feedback on the study instruments. The composition of the pilot sample was therefore more important than the size of the pilot sample. The evaluators selected for the pilot study were not only expected to possess the necessary expertise to engage critically with the study tasks but also to do so in a diligent manner. Table 13 presents the characteristics of the pilot sample.

Table 13

Pilot Sample Profile

Evaluator Profile	*n*
Current involvement in evaluation	
Execute evaluations	4
Academic interest in evaluation	4
Employment setting	
University	3
International donor agency	1
Highest academic qualification	
Master's degree	2
PhD	2
Academic Discipline	
Monitoring and evaluation	3
Environmental Science	1
Type of training in programme evaluation	
Self-educated	1
Master's degree	2
PhD	1
Number of evaluations completed in the last 5 years	
0-5	1
6-10	2
11-15	1
Practice context	
Developing countries	4

While most pilot participants (*n* = 3) were employed in a university setting, 75% had conducted more than six evaluations in the last five years and held at least a master's degree in programme evaluation. All participants also indicated that they were moderately to highly experienced in conducting outcome and impact evaluations. Based on the pilot sample's profile it is reasonable to assume that the

participants had the necessary expertise to engage critically with the pilot study tasks.

Data for the pilot study were collected electronically. A five point scale was used to capture participants' assessments. In line with Alkin and Christie (2005), pilot participants were also asked to respond to the following questions:

- What information did you think was missing from the scenarios?
- What major assumptions did you make in absence of that information?
- How similar (dissimilar) is the setting described in the scenarios to those in which you typically conduct evaluations?

It was important to extract this information as the design and presentation of the scenarios would inevitably affect how participants approach their assessment of evaluability.

A number of adjustments were made to the layout and manipulations embedded in each scenario based on the qualitative and quantitative data derived from the pilot study. There was a general consensus amongst pilot participants that the layout of the scenarios was not visually appealing. The scenarios, which initially consisted of multiple paragraphs of text, were long and difficult to read. Participants highlighted that some information was unnecessarily duplicated across the three scenarios. I therefore decided to condense each scenario by describing the evaluation context and presenting the instructions only once. The revised instructions page is presented next.

Instructions

We would like you to imagine that…

A funding agency wants to commission an outcome evaluation of an educational support program for high school students. The program is a one year, post-high school intervention for students who did not gain entry into a tertiary institution. The program offers a variety of academic activities. The client would like you to compare the performance scores of the beneficiaries with that of a comparison group who did not receive the program. You have had an initial meeting with the program staff and need to decide whether or not you will accept the evaluation contract.

You will be presented with three short scenarios of the program to be evaluated. Each scenario will be presented in a different colour (either blue, green or purple). Please read each scenario carefully and answer the three questions that follow. There are no predetermined right or wrong responses to these questions. **We request that you approach this exercise in the same manner that you would have if this were a real life situation.**

Each scenario will be presented only once. It might be useful to take down notes of what you think is important to consider in your decision making as you read each scenario.

Please click on the 'NEXT' button to proceed.

I also revised the layout of the scenarios. A question and answer format was used to organise the scenario descriptions into short segments of text. This revised format should allow participants to locate more easily information relevant to their decision making process. No changes were made to the evaluation context described in the scenarios as pilot participants indicated that the scenarios were highly realistic (*M* = 4.25) and similar to those in which they typically conduct evaluations (*M* = 4.50).

Five problematic manipulations and six problematic manipulations were identified in Scenario 1 and Scenario 2, respectively. Scenario 1 was designed to capture the following manipulations: (a) reliable programme data, (b) clearly defined service delivery, (c) inadequate budget, (d) difficulty in implementing the required methodology, and (e) difficulty in conducting the proposed evaluation. Most pilot

participants however failed to identify these intended manipulations. Table 14 presents the adjustments made to Scenario 1 and Scenario 2 in order to make the intended manipulations more salient. The same changes were replicated in Scenario 3, where applicable.

Table 14

Problematic Scenario Manipulations and Adjustments

Problematic manipulations	Adjustments
Scenario 1	
Reliable programme data	Specified that the evaluator will have access to *verified pre-programme and post programme data*
Clearly defined service delivery	Replaced the term service delivery by: *the manner in which the programme is delivered*
Inadequate budget	Replaced small budget by: *very tight budget*
Required evaluation methodology not feasible	Specified that evaluator will have to find a suitable matched comparison group
Type of evaluation required not feasible	Addressed by changes above
Scenario 2	
Unrealistic programme outcomes	Replaced improved career prospects by: *improved employment prospects after completion of tertiary studies*
Programme outcomes not measurable	Replaced general student development by: *socio emotional welfare*
Target beneficiaries not clearly defined	Deleted the term *socio economically disadvantaged* when describing the target beneficiaries in the instructions page to avoid contamination
Inadequate budget	Replaced small budget by: *very tight budget*
Required evaluation methodology not feasible	Specified that evaluator will have to find a suitable matched comparison group
Type of evaluation required not feasible	Addressed by changes above

Pilot participants responded negatively to the Q Sort task. They described the exercise as restrictive and cognitively overwhelming. Pilot participants were required to place a predetermined number of Q statements into five ranking categories (*Not at all important, Quite unimportant, Neither important nor unimportant, Quite important, and Essential*) based on the importance they assign to each evaluability criterion. For example, only three Q statements could be placed in the *Not at all important* and *Essential* categories. Figure 8 presents the forced distribution grid used in the pilot study.

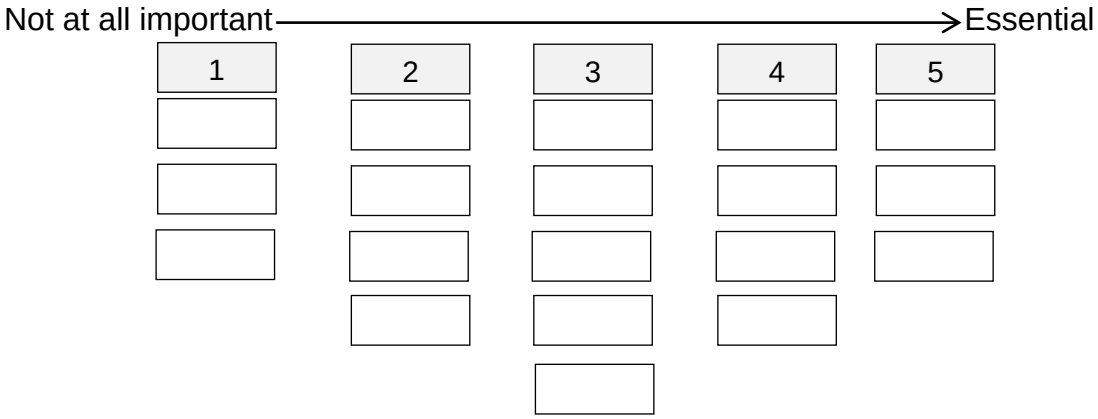


Figure 8. Forced Q Sort Distribution Grid.

The forced-choice condition of instruction largely contributed to the perceived difficulty and restrictive nature of the task. Most participants (*n* = 3) reported that they arbitrarily placed Q statements in the given categories in order to comply with the forced distribution requirement. This compromised the reliability of the Q Sort task. I addressed the highlighted issues by adopting a free-sort condition of instruction and a two-stage measurement protocol in the main study. This approach was expected to lower participants' frustration with the task and improve both the reliability and response rate of the study. While a free-sort condition of instruction will inevitably yield variations in the distribution of scores, this is not expected to affect the factor analytic solution (Brown, 1980; Thompson, 1998; Watts & Stenner, 2005).

Main Study

Evaluation practitioners from four countries (Brazil, South Africa, UK and USA) were the target population for the main study. Before recruiting prospective participants and administering the survey, the study materials for Brazil had to be translated into Portuguese by an independent translator. The Portuguese version was then back-translated into English by a second translator. The two versions were examined for linguistic congruence before they were logged on an online data collection platform called Qualtrics. An online data collection strategy was used as the sample of interest had a wide geographical distribution. Four independent survey links were generated, one for each sample of interest. The online data collection strategy was not expected to influence the results of the Q Sort task or scenario task. Van Tubergen and Olins (1979) and Reber, Kaufman and Cropp (2000), found the results of self-administered, electronic-based Q Sorts to be consistent with the traditional method of administration.

Data collection began once ethics clearance was obtained from the UCT Faculty of Commerce Ethics in Research Committee. Data were collected over a period of two months, starting early October, 2014. I took, on average, five weeks to collect sufficient responses from each cohort of interest. The participant recruitment strategy, the participant profile, the survey administration procedure, and the data analysis approach is described below.

Participant recruitment strategy.

I used a multistep participant recruitment strategy, which first involved identifying four professional evaluation associations, one in each country of interest, and enlisting their collaboration in the research. During this phase, I negotiated access to the membership database of the four participating associations (BMEN, SAMEA, UKES and AEA) and established the exclusion criteria: foreign affiliates, members with no practical experience in programme evaluation, and student members. While student members might have some experience in designing evaluations, I was concerned that they might be more inclined to reiterate their theoretical socialisation as opposed to engaging with the study tasks in the same manner that evaluators with actual

experience would. A purposive sampling strategy was therefore used to meet my sample requirements: Evaluation practitioners from Brazil, SA, UK and USA, with both specialist knowledge and experience in conducting evaluations.

The AEA provided me with a random listing of 1000 email addresses extracted from the association's membership database. Email addresses of student members and foreign affiliates were excluded from this listing. This random sample represented 13.7% of the total number of members affiliated to the association in 2014.

An invitation to participate in the study was sent to all 1000 email addresses in early October, 2014 via Qualtrics. The email delivery rate was 98.7%, with only 13 failed delivery messages recorded. Only 34 complete responses were received in the first 14 days, prompting me to extend the data collection deadline by two weeks. Two reminder emails were sent to the all 1000 potential participants, two weeks after the initial invitation was sent, and again one week later. Thirteen failed delivery messages were recorded on both occasions. An additional 32 complete responses were collected by late October from the 987 potential respondents.

The study was also advertised at the 28th Annual Conference AEA conference in Denver, Colorado (October 15-18, 2014). Thirteen potential participants, who were not on the initial email listing, were recruited at the conference. The survey link was forwarded to them via email, with a request to invite other eligible evaluation practitioners from their professional network to participate in the study. I extended the same invitation and request to at least 42 other prospective participants in my USA professional network. Twenty-one additional complete responses were collected from the USA cohort using this targeted strategy.

The other three participating professional associations (BMEN, SAMEA, and UKES) did not give me access to their membership database but collaborated in the following manner:

1) The BMEN administrator advertised the study on the BMEN website and disseminated the survey invitation via the association's mailing list in late October, 2014. The number of failed delivery messages was not recorded.

Fourty-nine complete responses were collected in the first week. A reminder email was sent in the second week via the same mailing list. The data collection deadline was extended by two weeks. Fourty-three additional complete responses were collected from the Brazil cohort by the end of the third week. I cannot accurately determine the number of responses (if any) that came from non-BMEN members.

2) The UKES administrator disseminated the survey invitation via the association's mailing list in late October, 2014. The number of failed delivery messages was not recorded. Only 10 complete responses were collected in the first week. No email reminder was sent. The UKES administrator re-advertised the study in the UKES e-bulletin one week after the initial study invitation was sent. The data collection deadline was extended by two weeks. No additional complete responses were collected over this time period.
3) The SAMEA administrator advertised the study on SAMEATalk, a moderated discussion and dissemination forum. Approximately 800 members subscribed to this forum in late October, 2014. Only six completed responses were collected in the first week. The study was re-advertised on the same forum one week later. The data collection deadline was extended by two weeks. No additional complete responses were collected over this time period.

I used a number of targeted strategies to address the low response rate from the SA cohort and UK cohort. As starting point, the survey invitation was sent directly to the 104 SAMEA members listed in the online membership directory in early November, 2014. The email delivery rate was 95.2%, with only five failed delivery messages recorded. A reminder email was sent a week later. Three additional completed responses was collected using this strategy.

The following combined strategies resulted in 31 additional completed responses from the SA cohort by mid-November, 2014.

1) I sent personalised emails to evaluation practitioners in my professional network, with a request to participate in the study and disseminate the survey invitation widely. The survey link was forwarded to at least 75 individuals with some involvement/interest in programme evaluation.

2) Evaluation practitioners employed by the following major evaluation/research consultancies were invited to participate in the study: Impact Consulting; Southern Hemisphere; Creative Consulting & Development Works; Benita Williams Evaluation Consultants; Khulisa Management Services; Mthente Research and Consulting Services; Evaluation Research Agency; and InsideOut M&E Specialists.
3) The study was advertised on the UCT Evaluation Group website. A notification was sent to the 56 members affiliated to this group in November 2014.
4) The UCT M&E alumni group, consisting of 40 graduates from the 2007 to 2013 M&E master's programme, was invited to participate in the study.

A similar approach was used to address the low response rate from the UK cohort. I capitalised on my professional network in the UK and invited evaluation practitioners from selected consultancies such as the Centre for Strategy and Evaluation Services (CSES), INTRAC, DMSS Research and Consultancy, and ITAD to participate in the study. The study was also advertised via the Monitoring and Evaluation NEWS mailing list available on www.mande.co.uk. While this membership list contained over 2000 subscribers in November 2014, not all of them were eligible to participate in the study. An additional 16 valid responses were collected from the UK cohort using this strategy, by the end of November, 2014.

The same incentive for participation (the opportunity to enter a lucky draw to win an IPAD) was advertised to all prospective participants, irrespective of the type of recruitment strategy used.

Participant profile.

Table 16 presents the estimated size of the target population, the total number of responses from each country of interest (realised sample), and the estimated overall response rate. It should be noted that is difficult to calculate the actual size of the target population and the actual response rate for three main reasons:

- I used a multistep participant recruitment strategy, whereby the initial sample of respondents recruited prospective participants from their professional network. A non-discriminative exponential version of the snowball method was used. While this method facilitated access to prospective participants who would have been otherwise difficult to reach, I cannot accurately determine the number of referrals made by the initial sample of respondents.

- It is difficult to define the boundaries and size of the target population as there are no formal certifications or licensure required to be an evaluation practitioner in Brazil, SA, UK or USA. I therefore cannot claim to have sampled a representative subset of evaluators from each country of interest.

- Most participants were recruited via the mailing lists of the participating associations. I did not have access to the mailing lists and therefore could not determine the actual size or accuracy of the mailing lists.

A rough estimate of the target population and the overall response rate could however be calculated using the approximate size of the mailing lists and the total number of survey invites sent directly by researcher. Table 15 presents the estimated target population, the realised sample and the estimated response rate for each country of interest.

Table 15

Estimated Target Population, Realised Sample, and Estimated Response Rate for Each Country of Interest

Country	Estimated Target population	Realised sample	Estimated Response rate
	N	*n*	%
Brazil	5000	197	3.9
South Africa	979	83	8.5
United Kingdom	512	81	15.8
United States of America	1055	143	13.6
Total	7546	504	6.7

Note. Only a small (undetermined) proportion of the BMEN mailing list represented eligible prospective participants.

An estimated overall response rate of 6.7% was achieved for the study. Of the 504 participants who accessed the study link, 245 completed the study in full, representing a 48.6% overall completion rate. The study response rate/completion rate is in line with similar simulation studies on evaluation practice (e.g., Azzam & Szanyi, 2011, with 212 completed responses from a random sample of 1500 AEA members, and a corresponding response rate of 14.1%).

Table 16 presents the study completion rate per country of interest.

Table 16

Study Completion Rate per Country of Interest

Country	Realised sample	No. of complete responses	Completion rate	Estimated response rate based on complete responses
	n		%	%
Brazil	197	92	46.7	1.8
South Africa	83	40	48.2	4.1
United Kingdom	81	26	32.1	5.1
United States of America	143	87	60.8	8.3
Total	504	245	48.6	3.2

It is clear from Table 16 that the USA had the highest study completion rate (60.8%) and UK had the lowest study completion rate (32.1%). It should be noted that even though the study had 245 complete responses, pairwise deletion was used for each statistical analysis to maximise the use of valid data. As such, the number of observations varied for each analysis, and in some instances exceeded 245. The handling of missing data in this study is described under the data analysis section.

After deleting individual cases with excessive levels of missing data and respondents with an academic interest in evaluation but no actual experience in conducting evaluations, the final sample for Brazil, SA, UK and the USA consisted of 91, 45, 30 and 94 respondents respectively. Table 17 presents the number of cases that were deleted from each dataset.

Table 17

Number of Deleted Cases and Final Sample per Country of Interest

Country	Cases with excessive levels of missing data	Cases with no evaluation experience	Final sample *n*
Brazil	92	13	91
South Africa	37	2	45
United Kingdom	51	0	30
United States of America	47	2	94
Total	227	17	260

It is clear from Table 17 that the Brazil dataset had the highest number of cases with excessive levels of missing data and respondents with no actual evaluation experience. The overall sample distribution is skewed, with UK and SA having the least number of respondents. If the Brazil and SA datasets are merged to represent evaluators from developing countries, the combined dataset would contain 136 valid cases. Similarly, if UK and USA datasets are merged to represent evaluators from developed countries, the combined dataset would contain 124 valid cases. These two distinct cohorts of evaluators are more or less of the same size, and hence provide me with the scope to perform meaningful comparative analyses.

Appendix E presents the defining characteristics of the final sample of 260 evaluation practitioners.

A cursory inspection of the data presented in Appendix E revealed that most participants in SA, UK and USA were currently involved in designing and conducting evaluations, amongst other evaluation related activities. The USA cohort had the highest percentage of participants (68.1%) employed in an evaluation job, followed by the SA (48.9%), UK (46.7%), and Brazil (11%) cohorts. Most participants from each cohort of interest were concentrated in particular employment settings, with 40.7% of the Brazil cohort employed in the public sector, 22.2% of the SA cohort working for non-governmental organizations, 26.7% of the UK cohort employed by private evaluation consultancies, and 20.2% of the USA cohort employed in a university setting.

Participants from the USA were highly experienced evaluators, with 28.7% having between 11 and 15 years of evaluation experience, and 30.9% holding a PhD in evaluation. Participants from the UK were also experienced evaluators, with 36.7% having between six and ten years of evaluation experience. Most of these participants (63.3%) however did not receive any formal training in evaluation. The Brazil and SA cohorts were the least experienced, with most participants (40.7% and 33.3% respectively) having between one and five years of evaluation experience. Most of these evaluators were either self-educated or completed a short course certificate in evaluation.

The overall sample consisted of highly active evaluators given that most participants from SA, UK and USA (63%, 70%, and 71.3% respectively) were currently working on an evaluation, and most participants from Brazil (44%) completed their last evaluation less than a month ago. Most participants also completed between one and five evaluations in the past five year, with the majority of participants from Brazil and SA (78% and 78.3% respectively) completing most of their evaluation work in developing countries, and the majority of participants from UK and USA (40% and 79.8% respectively) working mostly in developed countries.

Table 18 presents participants' self-rated level of experience in conducting different types of evaluations (a four-point scale ranging from *Not all experienced* to *Highly experienced* was used*)*.

Table 18

Self-rated Experience in Conducting Different Types of Evaluations

	Brazil *n*=78		SA *n*=45		UK *n*=30		USA *n*=86	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Evaluation readiness assessments	2.18	0.92	2.36	1.05	2.27	1.15	2.37	1.04
Needs assessments	2.54	0.95	2.42	0.97	2.31	1.05	2.76	0.92
Implementation/process evaluations	2.81	0.91	3.29	0.73	3.62	0.57	3.37	0.84
Outcome evaluations	3.06	0.83	3.19	0.81	3.65	0.63	3.45	0.71
Impact evaluations	2.40	0.96	2.54	0.99	3.27	0.87	3.07	0.84
Summative evaluations	2.19	1.02	2.97	0.96	3.00	1.02	3.20	0.91
Formative evaluations	2.56	1.03	3.11	0.92	3.00	1.13	3.20	0.89
Meta-analyses of evaluations	1.95	0.90	1.89	0.92	2.42	1.10	1.87	0.99

Participants across all four cohorts of interest reported that they were moderately to highly experienced in conducting outcome evaluations. This suggests that they had the expertise to engage with the simulated task, which called for an outcome evaluation. Participants were however not particularly experienced in conducting evaluation readiness assessments. It is important to note that lack of experience in a given area is not synonymous with lack of ability.

Survey administration procedure.

The survey link was included in the invitation email and participants could either click on the link or copy the URL in their internet browser to access the survey.

A decision was taken to standardise the presentation order of the study tasks. All participants were required to complete the scenario task first and Q Sort task second. This particular presentation order was chosen because exposure to the Q

statements might sensitise participants to respond in a particular way on the scenario task, thus contaminating their assessment of evaluability.

Respondents were, in the first instance, presented with a contextualised description of the scenario task, followed by specific instructions on how to approach the task. The three scenarios were then displayed in a randomised order, one at time, followed by a set of three identical items. Each scenario was presented in a different colour (either blue, green or purple) so that participants could differentiate between them. Participants were instructed to rate the evaluability of the programme depicted in each scenario on a 10-point scale, with *Evaluable with a lot of difficulty* and *Very easily evaluable* used as anchor points. Participants were then required to specify the three most important factors they considered in their assessments and indicate on a 10- point scale the likelihood that they would evaluate each programme. The scenario task was set up in such a way that participants could not modify their responses as they moved from one scenario to the other. After completing the scenario task, participants were provided with a list of 19 Q statements, presented in randomised order on the left of the screen. Participants were instructed to familiarise themselves with the type and range of statements presented before proceeding with the sorting task. They were then instructed to consider the importance they assign to each criterion when assessing the evaluability of a programme and drag each statement into the most appropriate box. Five boxes with the following labels were presented vertically on the right of the screen: *Not at all important, Quite unimportant, Neither important nor unimportant, Quite important, and Essential.* Participants could distribute any number of statements across the five boxes, and shift the statements from one box to the other until they were satisfied with their sorting.

Participants were then required to identify and rank-order the three criteria that they prioritised the most and three criteria that they prioritised the least when assessing the evaluability of a programme. They were instructed to work with the statements that they initially placed in the boxes labelled *Essential* and *Not at all important* to complete this exercise. This step was included in the protocol because participants might have placed all or most of the 19 Q statements in the *Essential* ranking category. Such a stance does not take into account the practical reality of an

evaluation context, whereby all desirable features are rarely present. By requiring participants to rank-order a set of evaluability criteria that they deemed essential, I attempted to re-create the reality in which they typically practice (i.e., one that inevitably requires them to make forced-choices at the implementation/practical level).

Some authors have argued that using a pre-determined Q set might restrict participants' responses. Watts and Stenner (2005, p. 78) dismissed the logic underlying this assumption, and contended that it "overlook[ed] the basic aims and premises of the method". To address the concerns of potential sceptics, participants in this study were required to specify any other evaluability criteria (not captured in the Q set) that they deemed essential.

After completing the Q Sort task, participants had to respond to 12 items (see Appendix F) relating to: their current involvement in evaluation, employment setting, highest academic qualification, type of training in evaluation, level of experience in conducting different types of evaluations, practice context, as well as, the number of evaluations that they conducted in the last five years.

Figure 9 summarises the presentation order of the various study tasks and the number of items embedded in each task.

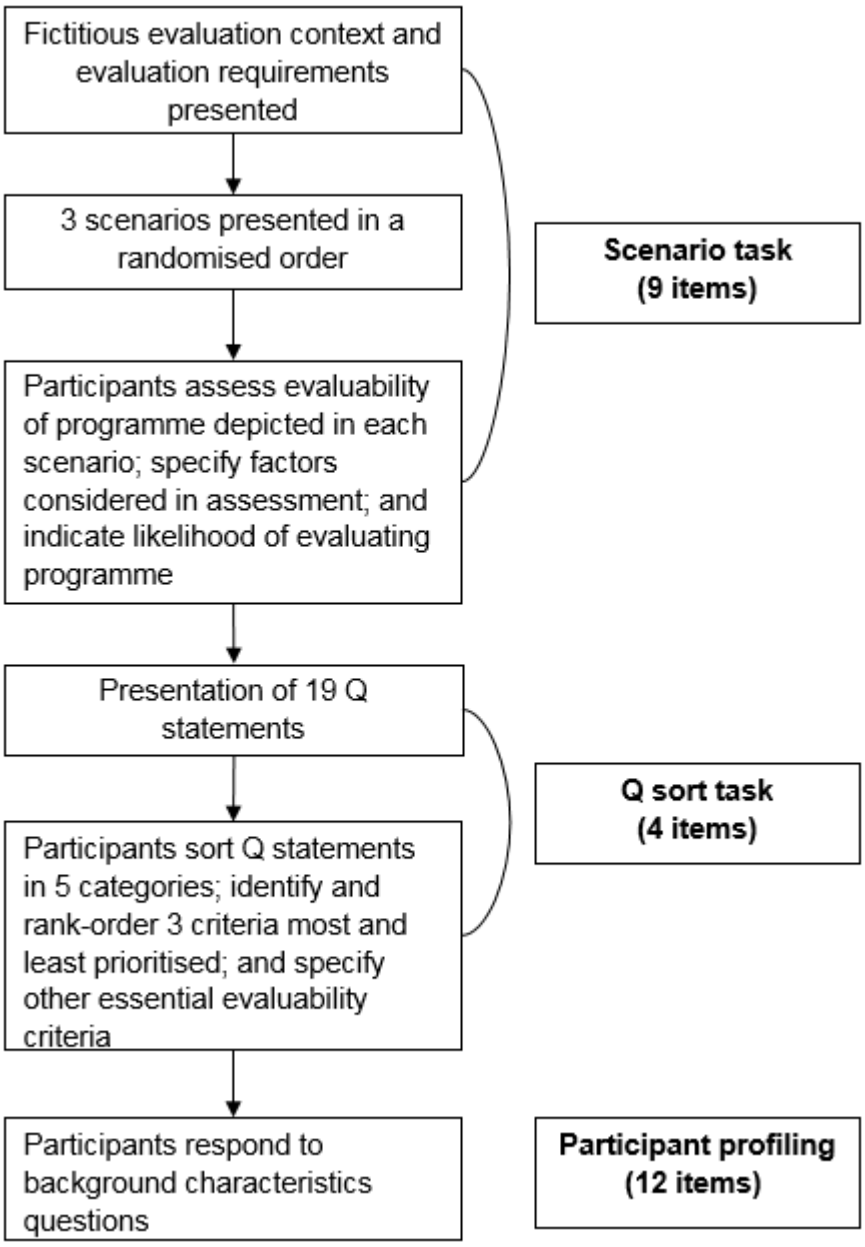


Figure 9. Survey Administration Procedure

A total of 25 items were embedded in the survey. While participants were instructed to complete the study in one sitting, a case-by-case inspection of the survey duration times revealed that a large proportion of participants did not do so. Forty-five participants completed the study over more than four hours. It is therefore difficult to accurately calculate the average time taken to complete the study in one sitting.

Data analysis.

Four separate datasets containing data for each of the four evaluator cohorts were generated from Qualtrics. The first steps in the data analysis process comprised of evaluating the extent of missing data, selecting the most appropriate multivariate techniques to analyse the data, and testing for the statistical assumptions underlying each technique. Each step and its corresponding outcome is described below.

Evaluation of missing data.

A four-step approach to evaluating missing data was used (Hair, Black, Babin, Tatham & Anderson, 2006). The first step involved the identification of any patterns that could characterise the nature of the missing data (e.g., systematic data entry error or nonresponse). It is reasonable to conclude that the missing data in the four data sets represent nonresponse by participants as the raw data were not entered manually. The second step involved an assessment of the extent of missing data for individual cases and variables across each of the four datasets. Appendix G presents the percentage of cases with valid and missing data on each numeric variable.

The extent of missing data varied across the 18 numeric variables (see Table G1 in Appendix G). The USA cohort had the lowest percentage of missing data across all numeric variables while the UK cohort had the highest percentage of missing data across all numeric variables (see Table G2 in Appendix G). While the extent of missing data was quite high (representing a high participant dropout rate), it was not concentrated in a particular set of variables, and hence can be assumed to operate in a random manner. Given that no specific non-random pattern was identified (step 3), complex imputation techniques, such as mean substitution, were not warranted. Instead, individual cases with excessive levels of missing data (less than 5 out of 18 variables completed) were deleted (step 4). The number of cases with valid data was sufficient for the selected analyses. Pairwise deletion was employed for each statistical analysis to maximise the use of valid data.

It is important to note that I cannot assume that the non-respondents or those who dropped out of the study were unlikely to have markedly different views on evaluability compared to those who had been successfully recruited and retained into the sample. I cannot rule out this possibility given the difficulty associated with investigating whether or not significant differences exist in the basic characteristics of respondents and non-respondents.

Selection of multivariate techniques.

I used the following data analysis techniques to answer the research questions:

- Q factor analysis (for research question 1).
- Correspondence analysis (for research question 2).
- Multinomial logistic regression (for research question 3).

The underlying principles of each technique are discussed below, followed by a systematic description of how I implemented each technique. The Statistical Package for the Social Sciences (SPSS) was used to perform the statistical analyses.

Q factor analysis.

Factor analysis is a technique used to analyse the structure of interrelationships among a set of variables. The aim is to condense the information contained in the original set of variables into a smaller set of composite dimensions (factors), with a minimum loss of information (Hair et al., 2006). A similar technique, known as the Q factor analysis, can be used to examine patterns of relationships among a set of respondents, with the aim of condensing the original set of respondents into distinctively different groups (factors) (Thompson, 1998; Du Plessis, 2005).In a Q factor analysis, the correlation matrix of individual respondents as opposed to variables are subject to factoring. What distinguishes the Q factor analysis from the conventional factor analysis is the organization of the raw data matrix, and not the mathematics of the factor analytic process (Thompson, 1998).

While a Q factor analysis determines which sets of respondents cluster together, it is different from a cluster analysis. Q factor analysis groups respondents based on their shared variance, while cluster analysis forms groupings/clusters using a distance-based similarity measure that calculates the degree of similarity between respondents' scores across a number of variables (Hair et al., 2006).

I adapted Hair et al.'s (2006) six-step factor analysis decision making process while implementing the Q factor analysis. I first specified the unit of analysis and organised the data matrix accordingly. The raw data in each of the four datasets were transposed so that the rows represented the Q statements and the columns represented the respondents' Q sorts. Ideally, the number of row replicates must be several times larger than the number of column entities to be factorised in order to extract more stable factors (Thompson, 1998). This was however not the case in the present study. Based on Eghbalighazijahani et al.'s (2013) findings, which supported the suitability of the Q method for both small and large P sets, and Field's (2013) assertion that cases-to-variables ratio make minimal difference to the stability of factor solution, this was not deemed to be an issue of concern.

I then calculated the input data for the analysis: a correlation matrix representing the degree of similarity/dissimilarity between respondents' Q sorts (Van Exel & de Graaf, 2005). The third step involved testing the assumptions underlying the Q factor analysis. Tests of normality, homoscedasticity, and linearity are rarely applied in the context of a factor analysis as, from a statistical standpoint, departures from these distributions only diminish the observed correlations (Hair et al., 2006). I therefore only assessed the factorability of the correlation matrix to justify the application of the Q factor analysis. Barlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, typically used for this purpose, were not deemed useful here as the small Q set to P set ratio would inevitably result in a non-positive definite matrix (Field, 2013; Hair et al., 2006). These tests can only be performed on a positive definite correlation matrix. I therefore visually examined the inter-correlation and communalities matrices to identify and exclude any factored entity with few correlations above .30 and/or a low proportion of common variance.

The fourth step involved deciding on the method of factor extraction (common factor analysis or components analysis) to be used and deciding on the number of factors to be selected to represent the underlying structure of the data. Although considerable debate remains over which factor model is more appropriate, in most applications, both component analysis and common factor analysis arrive at essentially identical results (Hair et al., 2006). Principle Component Analysis (PCA) is the default method of extraction in most statistical programmes. In this study, the PCA method was used because the objective was to summarise most of the factored entities (respondents) in a minimum number of factors, and then determine the substantive importance of each factor based on their corresponding eigenvalue. Different factors will represent different perspectives on evaluability, with respondents sharing a common perspective defining the same factor.

There is no exact quantitative basis for deciding the number of factors to be extracted (Hair et al., 2006). I used the following criteria to extract the most representative and parsimonious set of factors:

- Eigenvalue criterion: As per Kaiser's (1960) recommendation, only factors with eigenvalues greater than 1 were considered significant and retained for factor rotation. In other words, only factors that accounted for the variance of at least one factored entity were retained.
- Percentage of variance criterion: This criterion is met when a specified cumulative percentage of the total variance, extracted by successive factors, is achieved. The aim is to ensure that the derived factors explain at least a specified amount of variance and are practically significant. A threshold of 60% was considered satisfactory.
- Scree test criterion: A scree plot, which represents each eigenvalue plotted against the corresponding factor, was used to identify the optimum number of factors that can be extracted before the amount of unique variance begins to dominate the common variance structure. The point of inflexion was used as the cut-off point for retaining factors. Only factors to the left of the point of inflexion were retained (Field, 2013).

After extracting a set of factors, I proceeded with the fifth step, which involved simplifying the factor structure by means of a factor rotation. The aim was to facilitate the interpretation of the factor solution by minimising some of the ambiguities associated with initial unrotated factor solutions (Hair et al., 2006). While orthogonal rotational methods are more widely used than oblique methods, there are no compelling analytical reasons for favouring one method over the other. In line with Watts and Stenner's (2005) recommendation, I used the Varimax Rotation function to achieve a clearer separation of the factors and to increase the total variance explained. Because the factors were rotated orthogonally, the resulting factor matrix represented uncorrelated factors.

The factor matrix was then examined to identify significant factor loadings. The factor loadings indicate the extent to which each respondent's Q sort is associated with the rotated factors (Du Plessis, 2005). The highest loading for a particular respondent on any given factor was identified and its significance was determined (factor loadings > .50 were considered significant). Respondents who did not load significantly on any factors were deleted as their perspectives were idiosyncratic. Respondents with multiple significant cross loadings were also deleted as they did not possess well-defined and focused perspectives. My aim was to arrive at a factor solution with as many pure loadings as possible on a given factor. I then identified and deleted respondents who were not adequately accounted for by the factor solution (i.e., respondents with significant factor loadings but communalities < .50) before re-running the analysis, until a satisfactory factor solution with a high factor reliability is obtained. The composite reliability of a factor, which is an index of how much confidence can be placed in the factor, depends on the number of respondents that define it (Du Plessis, 2005). According to Brown (1993), five respondents per factor is sufficient to obtain a clear reading of the point of view that each factor characterises. In his view, any additional respondents would only marginally clarify the picture. Watts and Stenner (2005) proposed a less conservative requirement of at least two factor exemplars for a factor to be interpretable. Factor exemplars represent Q sorts that load significantly on only one factor.

I then examined the factor scores (Z-scores) computed as part of the analysis and labelled the final set of factors, which represent different perspectives on evaluability

captured in the form of Q sorts. A three-factor solution would mean that the discourse on evaluability (among the P set) breaks down into three distinct points of view. The factor scores indicate the importance of each Q statement in defining a given rotated factor. They represent the prototypical rankings of the Q statements in a given factor (Thompson, 1998) and facilitate cross-factor comparisons (Watts & Stenner, 2012). In line with Thompson (1998), a cut-off score of -1 or 1 was used to interpret the factor scores and identify the most important and least important evaluability criteria associated with a particular factor. It should be noted that a particular Q statement can contribute to more than one perspective.

As a final step, I interpreted the different factors/viewpoints that emerged from the Q analysis. Factor interpretation is an abductive process, which involves accounting for the entire item configuration captured in a given factor. In the absence of a set strategy for factor interpretation in the Q method literature, I applied the interpretative framework proposed by Watts and Stenner (2012) to arrive at a systematic and holistic interpretation of each factor. The first step involved the development of a crib sheet to facilitate the interpretative process. Q statements that fell under one of the following four categories were listed in the crib sheet based on the size and rank order of the factor scores: (a) statements given the highest ranking in each factor array, (b) statements given the lowest ranking in each factor array, (c) statements ranked higher in a given factor array compared to other factor arrays, and (d) statements ranked lower in a given factor array compared to other factor arrays. This approach ensured that the entire item configuration captured in a given factor array is taken into account during the interpretative process.

Once a crib sheet was compiled for each factor, I considered the positioning of each statement and applied the logic of abduction to explain its ranking and its significance in the context of the overall viewpoint. The aim was to generate iteratively the overall story underlying the various statement rankings and derive preliminary hypotheses that could account for a particular item configuration/factor array. In line with Watts and Stenner's (2012) recommendations, I considered "how things must feel for anybody who shares [a given] viewpoint" (p. 158) and the circumstances in which a particular statement might have been assigned a certain level importance by a group of evaluators. Watts and Stenner (2012) argue that a

good factor interpretation “should celebrate the first-person perspective and all the feelings that go with it” (p.159) and “must express what was impressed into [a particular factor] array” (p.163), given that the Q sort process is driven by the feelings and/or preferences of participants.

In order to clarify, support, or revise the preliminary account that was constructed, I examined the background characteristics of respondents who shared a common perspective. I also selected an appropriate label to represent each factor by examining the Q statements that distinguished them. While factor labelling is not a methodological requirement, it conveys in a parsimonious manner what distinguishes factors from one another. I chose a label that captured the general nature of each perspective. Statements with positive factor scores (i.e., characterised as essential) were given more weight in factor labelling. While this approach does not in any way capture the complexity of a given viewpoint, I strived to label each perspective in a manner that best integrates all the distinguishing statements associated with it.

The final interpretation of each factor is presented in narrative form. The relevant statements were linked together to create a unified account of the viewpoint embodied in each of the factors identified.

Correspondence analysis.

Correspondence analysis (CA) is a multivariate technique that allows researchers to analyse patterns of frequency-based associations within categorical data, in the absence of a priori expectations as to the nature of those associations (Doey & Kurta, 2011; Glynn, 2014). As in PCA, the aim is to reduce the dimensionality of a data matrix and detect underlying structures within the data (Nenadic & Greenacre, 2007). The major difference, however, is that PCA can only be performed on numerical data while CA can be used with categorical data (Hair et al., 2006).

CA is both a dimensional reduction and a perceptual mapping technique (Greenacre, 2010). With this technique, it is possible to generate a correspondence map that depicts, in a low dimensional space, the relative and simultaneous positioning of objects (i.e., any entity that can be evaluated in nonmetric terms) and variable

categories. More specifically, CA analyses two-way or multi-way contingency tables, which represent the cross-tabulation of categorical variables, and displays each row and column category as a point on the correspondence map by decomposing the total inertia (i.e., the variability) of the data table (Doey & Kurta, 2011). The simultaneous display of row and column data is unique to CA. Row and column categories with comparable patterns of counts/profiles will be positioned in relative proximity on the correspondence map. CA essentially converts the frequency of co-occurring categories into a metric measure of distance before plotting them in a low dimensional space (Glynn, 2014).

In the present study, the objects of interest were four different types of study tasks and the variable categories were 19 evaluability criteria. In the first three tasks, participants were required to respond to three different evaluation scenarios, each with a specific set of evaluability conditions. The fourth task was an acontextual exercise that required participants to sort a pre-determined set of evaluability criteria in order of importance. As such, the correspondence analysis was performed using four sets of data: the evaluability criterion that each respondent identified as *most important* in their separate assessments of Scenario 1, Scenario 2, and Scenario 3 (irrespective of whether the scenario was characterised as evaluable with ease or evaluable with difficulty); and the evaluability criterion that each respondent identified as *first on their priority list* in the Q Sort task. Table 19 presents the cross-tabulation of the objects and some variable categories of interest for illustrative purposes. In the actual analysis, the empty cells would capture the frequency with which each evaluability criterion was prioritised across the different study tasks.

Table 19

Cross-Tabulated Data: Type of Study Task by Evaluability Dimension

	Type of Study Task1			
Evaluability Criteria	Scenario 1	Scenario 2	Scenario 3	Q Sort
Clearly specified programme goals				
Implementation fidelity				
Willingness to collaborate				
Transparency about purpose				
Adequate budget				
Feasibility of implementing desired methodology				
Total				

In CA, a row or column profile represents the relative frequency of a set of observations in the contingency table. To calculate a profile value for each cell, the number of observations per row or column is added and then each observation is divided by the total. Because not all observations are of equal importance, CA uses weighted averages to compensate for this. The term mass refers to the weight of a given entry in the contingency table (Doey & Kurta, 2011). The weight for any entry can be calculated by dividing its value by *N* (i.e., the total for the table, which equals the sum of either the rows or columns). Inertia is the term used in CA to refer to the degree of variability in the contingency table. It can be calculated by dividing the total chi-square by the total of the frequency counts/observations (Hair et al., 2006). The higher the explained inertia, the better. When the inertia is high, row and column profiles have large deviations from their averages. The square root of the principal inertia represents the strength of association between row and column variables (Greenacre, 2007). The higher the inertia, the higher the row-column association. A total inertia value of above .20 is required for meaningful interpretation of the correspondence map.

It is important to note that in CA, the positioning of row and column points in the correspondence map is always relative (Greenacre, 2010). As such, Scenario 1 can

be positioned in close proximity to evaluability criterion 1, even if in absolute terms this criterion was prioritised more frequently in Scenario 2. In this particular case, the position of the points was determined by the relative prioritisation of the different evaluability criteria in Scenario 1. It is equally important to note that CA is an exploratory technique. I therefore cannot claim that the identified patterns are generalisable beyond the sample under investigation, or dismiss the possibility that they occurred by chance. Symmetric association between categorical variables should also not be confused with predictive association (Beh, Lombardo, & Simonetti, 2010).

The input data for the analysis consisted of qualitative responses independently coded by a trained research assistant (rater 2) using the coding scheme in Appendix H. The following types of responses were distinguished:

- Specific responses that were thematically in line with one of the evaluability dimensions / sub-dimensions.
- Generic responses with no underlying theme (e.g., required evaluation).
- Omissions (blank cells).
- Specific responses that were not thematically in line with any evaluability dimensions/sub-dimensions. These responses were coded as a separate category (Other).
- Double-barrelled responses. These responses were not coded.

Inter-rater reliability (based on a sample of 204 valid qualitative responses) was assessed using Cohen's Kappa. Results showed that the qualitative responses could be reliably distinguished. Inter-rater agreement was substantial, $\kappa = .81$.

Three variables were created in SPSS to represent the study tasks (four categories), each evaluability criterion (19 categories), and the associated frequencies. Each evaluability criterion was matched against each study task, thus creating 76 distinct entries in the SPSS file. The cases were first weighted by frequency before proceeding with the analysis. The variables *evaluability criterion* and *study task* were

inserted in the row and column profiles respectively. According to Doey and Kurta (2011), the positioning of variables on either axis does not affect the analysis.

Once the data were organised in a cross-tabulated form I checked that all values in the data matrix were positive. This is the only strict requirement of CA. Positive entries are required so that the distances between the points on the correspondence map are always positive (Doey & Kurta, 2011; Hair et al., 2006). It should be noted that CA does not make any distributional assumptions.

The CA output statistics relevant to the interpretation of the data included: the original contingency table, row/column profiles, association coefficients, chi-square test, principal inertias, and row/column coordinates. The data in the contingency table were examined to identify rare observations/objects and common observations/objects. Rare observations are often conceptualised as outliers in correspondence analysis (Greenacre, 2011). They can be identified by their high absolute coordinate values and their outlying position on a correspondence map.

While CA has been frequently criticised for being overly sensitive to rare observations, Greenacre (2011) has demonstrated empirically that these criticisms are unfounded and that the down-weighting or deletion of these outliers are not necessary in most cases. This is because the normalisation implied by the chi-square distance balances out the relative contributions of the rare and common observations, such that the more common observations dominate both the chi-square distances and the correspondence map. While rare observations do not necessarily influence the CA results, their outlying positions might make the remaining points cluster tightly together on the correspondence, thus making their interpretation more difficult (Alberti, 2013; Bendixen, 1996). If this were the case, I re-ran the analysis without the outlying points. If the relative positioning of the rows and columns remained virtually the same after the exclusion of potential outliers, I reverted to the original analysis. In line with Greenacre's (2011) recommendation, if a particular evaluability criterion was prioritised in only one study task and had a relatively low frequency (less than 5% of the total frequency on a given study task), it was removed from the analysis as it might represent an artifact. I anticipated such artifacts to be predominant in the Q Sort task given the nature of this task.

The first step in the interpretation of the analysis was to establish whether there was a significant dependency between the rows and columns in the contingency table (Bendixen, 1996). The higher the chi-square statistic, the higher the correspondence between rows and columns. The second step was to determine the dimensionality of the solution, that is, to specify how many dimensions will be used to represent graphically the dependency between row and column categories. The optimal number of dimensions is equal to the number of rows minus one or the number of columns minus one (if the number of columns is smaller the number of rows). As such, the optimal number of dimensions needed to represent 100% of the association between the categorical variables in this study would be three.

I specified a maximum of three dimensions to be extracted in the analysis but chose to represent graphically only two dimensions in order to facilitate the interpretation of the perceptual map. My aim was to retain relevant dimensions (i.e., only those that account for a significant proportion of the total inertia), generate a correspondence map that provides a good representation of the pattern of associations in the data, and arrive at a meaningful interpretation of the retained axes. A bi-plot/two-dimensional display was therefore derived in each instance. While a two-dimensional display might not transcribe all the inter/intra profile information, I prioritised ease of interpretation over completeness of description, in line with Greenacre's (1989) recommendations. The percentage of the total inertia was used to determine the accuracy/quality of the lower-dimensional projections (Greenacre, 2007). For example, if a display captured 97% of the inertia of the profiles, the loss of information (residual inertia or error) would be minimal.

The third step involved interpreting the two-dimensional map. I used the clustering approach as opposed to the factor analytic approach to do so. In the factor analytic approach, dimensions or axes are first interpreted, followed by an interpretation of the points with respect to these axes. The clustering approach, on the other hand, concentrates directly on the distance between points on the map (Kennedy, Riquier, & Sharp, 1996).

Two types of comparisons were possible using this approach: between categories of the same row or columns, and/or between row categories and column categories. I

was mainly interested in column to column comparisons (i.e., how a particular study task compared to other study tasks) and row to column comparisons. While there are some debates on the appropriateness of directly comparing row and column categories in symmetric displays, general comparisons can be made if a standardisation technique is applied (Doey & Kurta, 2011; Hair et al., 2006). In a symmetric display, the separate configurations of row and column profiles are superimposed on a joint map, and as such, the distance between row and column categories cannot be interpreted meaningfully (Greenacre, 2007). I used symmetrical normalisation (an option available in SPSS) to standardise the row and column data.

I used the following guidelines adapted from Alberti (2013); Doey and Kurta (2011); Hair et al., (2006); Hoffman and De Leeuw (1992); and Yelland (2010) to interpret the two-dimensional displays:

- The axes of the low-dimensional display are called principal axes. The horizontal axis represents the first dimension/principal axis and accounts for most of the inertia. The vertical axis represents the second dimension/principal axis and explains the second largest percentage of the inertia. A descriptive name can be assigned to each axis based on the positioning of the points and knowledge of which row and/or column categories have contributed the most to each axis.
- The distance between two X points or two Y points is related to the homogeneity of their profiles. As such, evaluability criteria that have been prioritised in the same study tasks will tend to be close, and study tasks sharing the same evaluability criteria will also tend to be close on the correspondence map.
- The origin of the axes represents the centroid (i.e., the average profile).
- A category point with low marginal frequency will be plotted towards the edge of the map, while a category point with high marginal frequency will be plotted nearer to the origin of the map. As such, evaluability criteria with profiles similar to the average profile will be plotted more towards the origin, while evaluability criteria with unique/unusual profiles will appear near the edges.

- The average row profile and the average column profile are situated at the origin. Row and column points that are close to the average profile (i.e., close to zero) contribute minimally to the inertia explained by the principal axes. As such, points that lie between -.2 and +.2 on a particular dimension were considered not to show any meaningful associations on that particular dimension.
- Distances are interpreted separately on each dimension. Two points might be in close proximity on one dimension but are not necessarily close together on the other dimension. Points that are close together on both dimensions are considered to correspond better than points that are close together on a single dimension. I focused on points that were close together on both dimensions.
- Row-to-column distances are used to assess the correspondence between row and column categories. Each evaluability criterion will lie more or less in the direction of or closer to the study task in which the evaluability criterion's profile is prominent.

In line with Yelland (2010) the following lines were drawn on the biplot to interpret the row-to-column distances:

1. Task-centroid lines (lines drawn from the centroid to each study task point).
2. Criterion-centroid lines (lines drawn from the centroid to each criterion point).
3. Criterion-task perpendiculars (perpendicular lines drawn from each criterion point to each task-centroid line).

The angle between a criterion-centroid line and a task-centroid line represents the strength of the association between the criterion and the study task from which the lines were drawn. In particular acute angles indicate a positive association while obtuse angles indicate a negative association between these points (Pusha, Gudi, & Noronha, 2009). However, the angle between a criterion-centroid line and a task-centroid line does not show the relative frequency/prioritisation of a given criterion in

a study task compared to other study tasks. The point of intersection of the criterion-task perpendicular with the task-centroid line indicates the relative frequency/prioritisation of a given criterion in a study task compared to other study tasks. When the angle is acute, the further this point of intersection is from the origin, the higher the relative frequency of the criterion in the task. Figure 10 illustrates the relevant lines used in the interpretation of row-to-column distances in CA biplots (data from the USA cohort was used to derive this map).

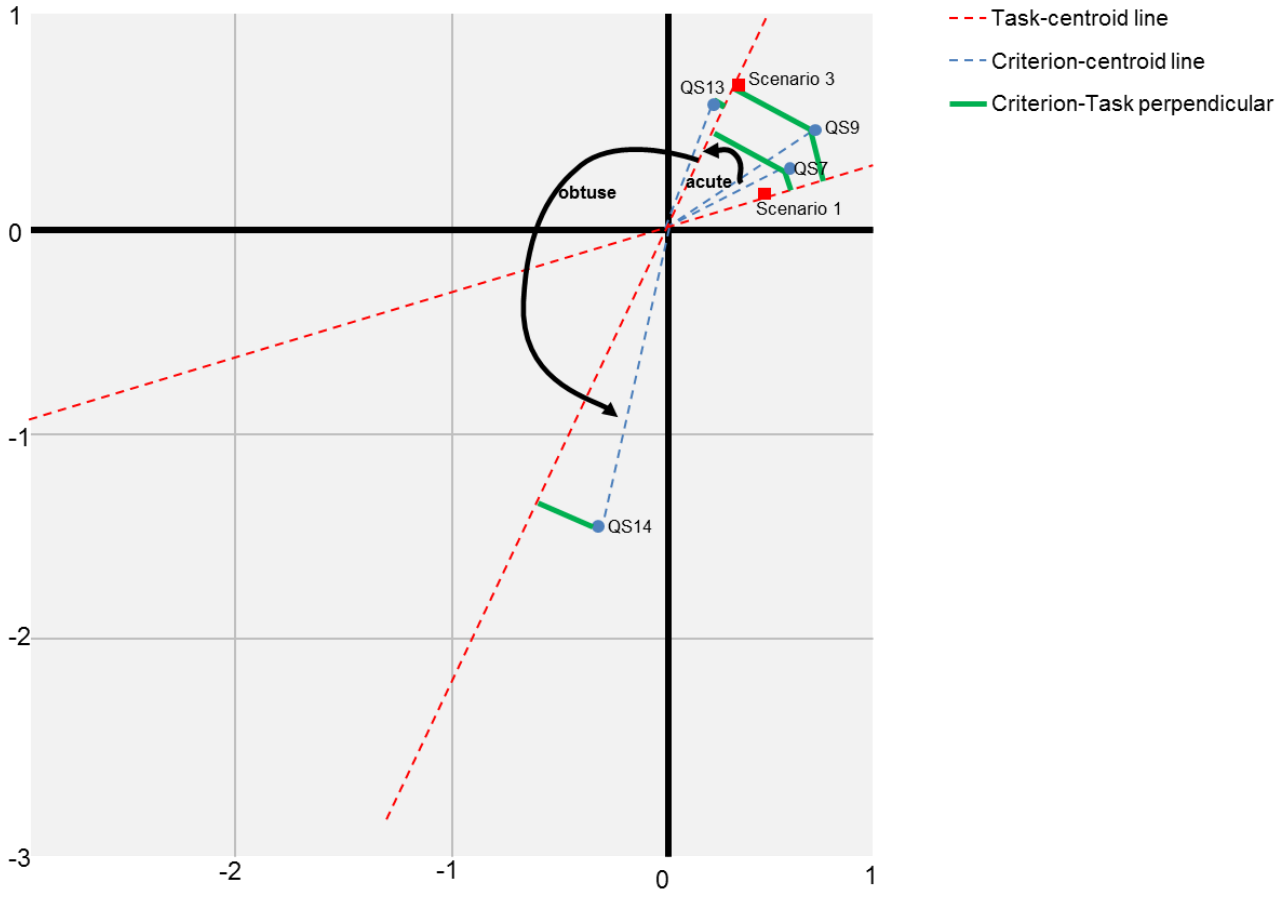


Figure 10. Method for interpreting row-to-column distances in CA biplots

Figure 10 depicts the results for Scenario 1 (S1) and Scenario 3 (S3). In the upper right quadrant, the perpendicular lines drawn from QS9 intersect the S1-centroid and S3-centroid lines the furthest away from the origin. The angle between the QS9-centroid line and the S1-centroid and S3-centroid lines is acute, indicating that QS9 is prioritised more frequently in these two study tasks relative to other study tasks.

In the lower left quadrant, the perpendicular line drawn from QS14 intersects the S3-centroid line the furthest away from the origin. The angle between the QS14-centroid line and S3-centroid line is however obtuse, indicating that QS14 is prioritised less frequently in Scenario 3 relative to other study tasks.

Multinomial logistic regression.

Multinomial logistic regression (MLR) is an extension of binary logistic regression, a statistical technique used to predict a two-category outcome variable from a set of predictor variables (Burns & Burns, 2008). MLR, on the other hand, can be applied when the dependent variable (DV) has more than two ordered or unordered categories. Both categorical and continuous independent variables (IVs) can be used as predictors in MLR, a feature that distinguishes MLR from discriminant analysis. In MLR odds ratios are used as estimators for the predictor variables and one of the categories of the DV is specified as a reference category, against which all the other categories are compared (Petrucci, 2009). The choice of reference category drives the interpretation of the results.

The assumptions of MLR are identical to those of binary logistic regression, with one exception: the outcome variable has to follow a multinomial rather than binomial distribution. Like binary logistic regression, MLR does not assume a linear relationship between the DV and IVs, or that these variables are normally distributed (Burns & Burns, 2008). MLR is often chosen over discriminant analysis because it does not impose strict data requirements. MLR does, however, require careful consideration of sample size, exclusion of outliers, and absence of multicollinearity. It also assumes that the categories of the DV are independent and mutually exclusive. A minimum of 10:1 case-to-variable ratio is recommended for the MLR analysis to be stable with 20:1 ratio being optimal (Petrucci, 2009).

MLR was used in this study to predict the evaluability decisions of evaluators based on their profile. More specifically, the IVs of interest were: level of experience, and context of practice. To comply with sample size requirements, I collapsed the initial five categories of experience into three categories representing low, medium and

high level of experience. I decided to retain the three categories of practice context as these could not be meaningfully collapsed into fewer categories.

Table 20 presents the predictors of interest, their corresponding categories, the description of the collapsed categories (where relevant), and the number of valid responses per category. It is clear that the two categorical IVs are conceptually different. I can therefore reasonably assume that they would not be strongly correlated.

Table 20

Study Predictors, Category Description and Total Number of Valid Responses

IV	Categories	Description	*n*	Valid %
Level of experience	Low	≤ 1 year to 5 years	100	43.9
	Medium	6 to 10 years	51	22.4
	High	11 to 15 years	77	33.8
Practice context	Developed countries	-	123	53.9
	Developing countries	-	90	39.5
	Both developed and developing countries	-	15	6.6

Note: Ratio of cases to levels of all variables= 15.2 (228/15), suggesting adequate sample size.

The three DVs were assessment of evaluability, likelihood of conducting an evaluation, and prioritisation of evaluability criteria. Participants' assessments of evaluability were collapsed into three categories corresponding to the level of difficulty associated with evaluating the programme depicted in each of the three scenarios. Likelihood of conducting an evaluation, given the specifics of each scenario, was categorised as either low, medium, or high. The evaluability criterion prioritised in each scenario was coded under one of the following three categories: programme structural features, stakeholder characteristics, and logical requirements. Table 21 presents the dependent variables, their corresponding categories, the description of the collapsed categories, and the number of valid responses per category

Table 21

Independent Variables, Associated Categories, Descriptions, and Valid Responses

IV	Categories	Description	Scenario 1	Valid	Scenario 2	Valid	Scenario 3	Valid
			n	%	*n*	%	*n*	%
Difficulty level of evaluation	Low	Score of ≥ 7	83	32.3	32	12.5	12	4.8
	Medium	Score of 5 or 6	55	21.4	46	17.9	39	15.5
	High	Score of ≤4	119	46.3	179	69.6	209	79.8
Likelihood of conducting evaluation	Low	Score of ≤4	108	42.0	153	60.2	161	64.4
	Medium	Score of 5 or 6	61	23.7	56	22.0	58	23.2
	High	Score of ≥ 7	88	34.2	45	17.7	31	12.4
Prioritisation of evaluability criteria	Programme Features	Criteria 1- 12	103	48.6	92	45.5	94	51.6
	Stakeholder Characteristics	Criteria 13-15	56	26.4	45	22.3	50	27.5
	Logistical Requirements	Criteria 16-19	53	25.0	65	32.2	38	20.9

I merged the data collected from each evaluator cohort into a single dataset to increase the power of the regression model. The MLR analysis was performed in SPSS. The IVs were simultaneously entered into the program as factors, with one DV at a time (the MLR was run six times in total). The minimum specification for the analysis is one DV and one factor or covariate predictor (SPSS Inc., 2006). The default SPSS reference category was used given that the last category of each DV had the highest frequency. The main effects model was selected as I was interested in investigating the main effects before exploring any two-way interactions. The following statistics/tables were generated for analysis: (a) case processing summaries (representing the frequencies for the IVs and DVs retained in the analysis), (b) pseudo R-squares, (c) model fitting information, (d) classification tables, (e) likelihood ratio tests, and (f) parameter estimates. I then proceeded with the following six-step analysis process:

1. Detecting multicollinearity.

Before interpreting the MLR results, I checked the standard errors associated with the *B* coefficients in the parameter estimates table. A standard error greater than two is indicative of high multicollinearity between the IVs (Petrucci, 2009). While multicollinearity does not change the estimates of the parameters, it affects the reliability of the results.

2. Assessing model fit and the overall relationship between the IVs and DV.

Model fit in MLR can be assessed in a number of ways. I first examined the model fitting information table and compared the -2 log likelihood values for the intercept only (i.e., a model that does not include any predictor variable), and the final model that includes the specified IVs. This difference follows a chi-square distribution, and is referred to as the model chi-square (Bayaga, 2010).The model chi-square statistic and the associated test for statistical significance were used to determine if there was a significant improvement in the model after adding the IVs. A *p* value of less than .05 would indicate significant improvement in model estimation and the presence of a relationship between the outcome variable and predictor variables.

I then examined the Pearson chi-square and deviance statistics presented in the goodness of fit table to assess overall model fit. Here, statistical significance was not desired as it would indicate that the model does not fit the data well (Petrucci, 2009).

3. Assessing the strength of the relationship.

While SPSS generates pseudo R square measures to estimate the strength of an established relationship, these measures do not estimate accuracy or errors associated with the model (Bayaga, 2010). The Cox and Snell R square and the Nagelkerke R square values can however provide a cursory indication of the amount of variation in the DV. It should be noted that while the Cox and Snell R square takes into account sample size, it cannot achieve a maximum value of 1.

4. Assessing the contribution of each IV to the model.

The output of the likelihood ratio test was used to determine whether or not the main effect of each predictor variable is significant. It should be noted that this test does not indicate whether or not a given IV is significant in distinguishing the different categories of the DV (Bayaga, 2010).

5. Interpreting coefficients.

The parameter estimates table was examined next. The relevant statistics for all paired categories of the dependent variable, including the exponentiated value of the estimated *B* coefficient, and significance values associated with the Wald statistic were captured in this table. The Wald test evaluates whether or not the IV is significant in distinguishing the different categories of the DV. The estimates presented are relative to the reference category specified in each analysis.

In line with Hair et al. (2006), exponentiated coefficients/odds ratios greater than 1 were interpreted as indicating a positive relationship/greater likelihood for the outcome of interest, while values less than 1 represented a negative relationship/ lower likelihood for the outcome of interest. An exponentiated coefficient of 1

denotes no effect (i.e., unit changes in the predictor variable does not affect the outcome variable).

If a particular predictor was not significant, the model was re-run after excluding that variable. I decided to use a more conservative significance level to account for the increased Type I error resulting from the large number of statistical tests being run. A corrected *p* value of .025 (standard *p* value of .05 divided by the total number of predictor variables) was used.

6. Evaluating the usefulness of the model.

I compared the overall classification accuracy rate (from the classification table) with the proportional by chance accuracy rate to assess the usefulness of the final model. The proportional by chance accuracy rate is calculated by summing the squared percentage of cases in each category. For a model to be useful, there must be a 25 % improvement over the rate of accuracy achievable by chance alone (Hair et al., 2006).

Summary

The primary purpose of the present study was to investigate whether evaluators from four different countries share a common and consistent perspective towards evaluability. The second purpose was to isolate the defining characteristics of the evaluators who shared similar perspectives, and investigate whether their level of experience, and practice context predict their actions and choices. In this chapter the method used to develop and pilot the study instruments (three scenarios and a Q Sort task); the procedure used to administer them to a purposive sample of evaluators from Brazil, SA, UK, and USA; and the statistical analyses used to analyse the data collected were discussed. The rationale for using a particular data collection strategy, participant recruitment approach, and statistical technique was provided. The main statistical analyses consisted of: Q factor analysis for research question 1, correspondence analysis for research question 2, and multinomial logistic regression for research question 3.

CHAPTER FIVE

RESULTS

This chapter presents the results of the statistical analyses (Q factor analysis, correspondence analysis and multinomial logistic regression) used in this study. Pairwise deletion was used for each statistical analysis to maximise the use of valid data. The first two statistical analyses were performed on the dataset of each evaluator cohort (i.e., the analyses were replicated across all four datasets), while the third analysis was performed on a single dataset, consisting of data points from all evaluator cohorts.

A step by step description of how each analysis was conducted is presented in Chapter 4. For the sake of brevity and to avoid repetition and tedium, only the main results of the study (i.e., those that address the research questions directly) are presented in this chapter. Supplementary results, such as descriptive statistics and results related to assumption testing and statistical manipulations associated with each analysis, are reported in Appendices I and K. These include: (a) factor extraction and rotation, (b) deletion of Q respondents with problematic factor loadings, (c) identification and deletion of outliers in the correspondence analysis, and (d) calculation of the overall classification accuracy rate and proportional by chance accuracy rate for the regression analysis.

The main results are organised under the corresponding research question:

Do Evaluators Share a Common Perspective Towards Evaluability? If Not, What Perspectives Can be Empirically Identified and What Evaluator Types are Most Associated With these Perspectives?

Q factor analysis was used to explore whether evaluators shared a unified or divergent perspective towards evaluability. I was primarily interested in the identification of dominant/distinctive perspectives that may be unique to different cohorts of evaluators, the specific criteria that characterise those perspectives, and

the profile of evaluators most associated with these perspectives. Relevant descriptive statistics are presented in Tables I1-I2 in Appendix I.

At least two major perspectives with high factor reliability were retained for each evaluator cohort (see Table 22 below and Tables I9-I12 in Appendix I for final rotated solutions)

Table 22

Number of Factors Retained per Evaluator Cohort and Percentage of Variance Explained by each Factor

Evaluator cohort	*n*	No. of factors with high factor reliability	% of variance explained by each factor	Cumulative %
USA	56	4	Factor 1: 24.5% Factor 2: 23.2% Factor 3: 8.9% Factor 4: 8.8%	65.5%
UK	13	2	Factor 1: 31.1% Factor 2: 25.0%	56.1%
Brazil	36	4	Factor 1: 22.3% Factor 2: 14.9% Factor 3: 13.0% Factor 4: 10.4%	60.7%
South Africa	16	2	Factor 1: 27.8% Factor 2: 16.6%	44.4%

In line with Bruce (1998), a cut-off score of (-) 1 was used to interpret the factor scores and identify the most important and least important evaluability criteria associated with a particular factor. The factor scores are in z-score form (i.e., they have a mean of 0 and a standard deviation of 1). The results for each evaluator cohort are presented next.

USA cohort: Evaluability perspectives and associated evaluator types.

The Q statements most useful in defining Factor 1 were QS13 (-1.3), QS17 (-1.3), QS18 (-1.3), QS19 (-1.1), QS12 (1.4), and QS8 (1.7). While the scaling direction used in the Q Sort task was arbitrary, with 0 representing the *Not at all important* category and 4 representing the *Essential* category, it needs to be considered in the interpretation process. As such, the 22 respondents most associated with Factor 1 considered the following two evaluability criteria as essential: *programme theory is explicitly stated* (QS8) and *programme is implemented as intended* (QS12). The four evaluability criteria not considered important at all were: *stakeholders are willing to collaborate with the evaluator* (QS 13), *timeframe is adequate to complete the evaluation* (QS17), *type of evaluation required (process, outcome or impact) is feasible* (QS18), and *required evaluation methodology is feasible* (QS19).

I applied the interpretative framework proposed by Watts and Stenner (2012) to arrive at a systematic and holistic interpretation of Factor 1 and subsequent factors (see Chapter 3 for a description of this framework). Factor 1 reflects a perspective that favours an explicit change logic and implementation fidelity but minimises the importance of logistical imperatives and stakeholder collaboration, authority, and transparency (see Table 23). It would seem that the underlying focus is on opening the black box of evaluation and making the underlying assumptions and implementation of the programme activities clear. This could be construed as the essence of a theory-driven evaluation approach. I therefore labelled this perspective as *theory-driven*.

Table 23

Factor 1 Crib Sheet: USA Cohort

Criteria	Q Statement	Factor Score
Items with the highest rankings in Factor 1 array	Programme theory is explicitly stated (QS8)	1.7
	Programme is implemented as intended (QS12)	1.4
Items ranked higher in Factor 1 array than any other factor array	Programme goals are clearly specified (QS1)	1.3
	Programme theory is explicitly stated (QS8)	1.7
	The manner in which the programme is delivered is clearly defined (QS10)	0.8
	Target beneficiaries are clearly defined (QS11)	0.6
	Programme is implemented as intended (QS12)	1.4
Items ranked lower in Factor 1 array than any other factor array	Stakeholders are willing to collaborate with the evaluator (QS13)	-1.3
	Stakeholders have authority to act on evaluation findings (QS14)	-0.8
	Stakeholders are transparent about the purpose of the evaluation (QS15)	-1.0
	Budget is adequate for the evaluation (QS16)	-1.0
	Timeframe is adequate to complete the evaluation (QS17)	-1.3
	Type of evaluation required is feasible (QS18)	-1.3
	Required evaluation methodology is feasible (QS19)	-1.1
Items with the lowest rankings in Factor 1 array	Stakeholders are willing to collaborate with the evaluator (QS 13)	-1.3
	Timeframe is adequate to complete the evaluation (QS17)	-1.3
	Type of evaluation required (process, outcome or impact) is feasible (QS18)	-1.3
	Required evaluation methodology is feasible (QS19)	-1.1

Factor 2 seems to reflect evaluators' concern with mechanisms that support the utilisation of results by intended users (stakeholder transparency, authority and consensus) as opposed to the quality and accessibility of evaluation data (see Table 24). I therefore labelled this perspective as *utilisation-focused*. Twenty-two respondents were most associated with Factor 2.

Table 24

Factor 2 Crib Sheet: USA Cohort

Criteria	Q Statements	Factor Score
Items with the highest rankings in Factor 2 array	Stakeholders are transparent about the purpose of the evaluation (QS15)	2.3
	Stakeholders have authority to act on evaluation findings (QS14)	1.8
	Programme outcomes are realistic (QS2)	1.1
Items ranked higher in Factor 2 array than any other factor array	Stakeholders agree on programme goals (QS4)	0.9
	Stakeholders have authority to act on evaluation findings (QS14)	1.8
	Stakeholders are transparent about the purpose of the evaluation (QS15)	2.3
Items ranked lower in Factor 2 array than any other factor array	Programme outcomes are measurable (QS3)	-1.3
	Programme data are adequate (QS5)	-1.1
	Programme data are reliable (QS6)	-1.0
	Programme data are easily accessible(QS7)	-0.6
Items with the lowest rankings in Factor 2 array	Programme outcomes are measurable (QS3)	-1.3
	Programme data are adequate (QS5)	-1.1
	Programme data are reliable (QS6)	-1.0

The interpretation of Factor 3 and Factor 4 was attempted for exploratory purposes, keeping in mind that due to the relatively low proportion of variance explained by these two factors (8.9% and 8.8% respectively), and the small number of respondents defining each factor (n = 6), it might be difficult to draw concrete

conclusions. The item configuration of both Factor 3 and Factor 4 is not related to any explicit notion of evaluation practice (see Tables I13-I14 in Appendix I), thus making the characterisation of these perspectives problematic. For example, it is difficult to reconcile in a meaningful manner specific aspects of programme theory (plausibility and articulation), data collection (accessibility) and logistical requirements (budget adequacy and feasibility of conducting desired type of evaluation), and use these to formulate a well-integrated approach to assessing programme evaluability. A decision was taken not to retain these two factors for further analysis and discussion.

Only two well-defined perspectives were therefore empirically identified for the USA cohort. The profile of evaluators who defined each perspective was somewhat different, particularly in terms of employment setting, and level of training and experience in evaluation (see Table I15 in Appendix I). Most evaluators (*n* = 13) who adopted a theory-driven approach (cohort 1) either worked in the public sector or as independent evaluation consultants, and had either a Master's degree or a PhD in evaluation. Seventy-two point seven percent (72.7%; *n* = 16) of evaluators in this cohort had at least six years of experience in conducting evaluation, with the majority (*n* = 11) having between 11 to 15 years of experience, and above.

Most evaluators (*n* = 12) who adopted a utilisation-focused approach (cohort 2) were employed in either a university or non-profit setting. Their training was limited to self-education or a short course certificate in evaluation. While some evaluators (*n* = 7) in cohort 2 had over 15 years of experience in evaluation, 50% (*n* = 11) had only between one and five years of experience in the field.

UK cohort: Evaluability perspectives and associated evaluator types.

The first factor extracted for the UK cohort reflects a perspective that minimises the importance of stakeholder collaboration, authority, and transparency, and certain logistical imperatives (see Table 25). It would seem that the underlying focus is on the ability to measure implementation fidelity, and explain why a programme worked or did not work (a plausible theory and clearly defined target beneficiaries can facilitate this task). This perspective is similar to the first factor identified for the USA

cohort, in terms of overall focus. As such, I decided to label it as *theory-driven*. Eight respondents were most associated with this perspective.

Table 25

Factor 1 Crib Sheet: UK Cohort

Criteria	Q Statement	Factor Score
Items with the highest rankings in Factor 1 array	Programme is implemented as intended (QS 12)	2.5
Items ranked higher in Factor 1 array than any other factor array	Stakeholders agree on programme goals (QS4)	0.8
	Programme data are adequate (QS5)	0.6
	Programme data are easily accessible (QS7)	0.1
	Programme theory is plausible (QS9)	0.8
	Target beneficiaries are clearly defined (QS11)	0.7
	Programme is implemented as intended (QS 12)	2.5
Items ranked lower in Factor 1 array than any other factor array	Programme outcomes are realistic (QS2)	-0.2
	Stakeholders are willing to collaborate with the evaluator (QS13)	-0.8
	Stakeholders have authority to act on evaluation findings (QS14)	-1.3
	Stakeholders are transparent about the purpose of the evaluation (QS15)	-1.2
	Type of evaluation required is feasible (QS18)	-1.7
Items with the lowest rankings in Factor 1 array	Type of evaluation required is feasible (QS18)	-1.7
	Required evaluation methodology is feasible (QS19)	-1.3
	Stakeholders have authority to act on evaluation findings (QS14)	-1.3
	Stakeholders are transparent about the purpose of the evaluation (QS15).	-1.2

The second factor reflects a perspective that emphasises the need for: (a) a logic model that articulates realistic and measurable outcomes, and (b) stakeholders' transparency and authority. An explicit programme delivery plan and certain logistical requirements are not of high priority (see Table 26). There seems to be a dual focus on programme theory, and the necessary conditions for utilisation of evaluation findings. This factor reflects a combined *theory-driven and utilisation-focused perspective,* and was therefore labelled as such. Five respondents were most associated with this perspective.

Table 26

Factor 2 Crib Sheet: UK Cohort

Criteria	Q Statement	Factor Score
Items with the highest rankings in Factor 2 array	Stakeholders are transparent about the purpose of the evaluation (QS15)	2.2
	Stakeholders have authority to act on evaluation findings (QS14)	1.7
	Programme outcomes are realistic (QS2)	1.2
Items ranked higher in Factor 2 array than any other factor array	Programme outcomes are realistic (QS2)	1.2
	Programme outcomes are measurable (QS3)	0.7
	Programme theory is explicitly stated (QS8)	0.7
	Stakeholders have authority to act on evaluation findings (QS14)	1.7
	Stakeholders are transparent about the purpose of the evaluation (QS15)	2.2
Items ranked lower in Factor 2 array than any other factor array	The manner in which the programme is delivered is clearly defined (QS10)	-1.1
	Budget is adequate for the evaluation (QS16)	-1.0
	Required evaluation methodology is feasible (QS19)	-1.4
Items with the lowest rankings in Factor 2 array	Type of evaluation required is feasible (QS18)	-1.5
	Required evaluation methodology is feasible (QS19)	-1.4
	The manner in which the programme is delivered is clearly defined (QS10)	-1.1
	Budget is adequate for the evaluation (QS16)	-1.0

There were no striking differences between the profiles of evaluators who use a theory-driven approach to assess the evaluability of a programme and those who use a combined theory and utilisation-focused approach (see Table I16 in Appendix

I). The small number of evaluators subscribing to each perspective made it difficult to identify any discernible patterns in this case.

Brazil cohort: Evaluability perspectives and associated evaluator types.

Fourteen respondents were most associated with the first factor extracted for the Brazil cohort. This perspective emphasises the need for an explicit and plausible theory of change, which operationalises clearly specified and agreed-upon programme goals. Evidence that the programme has been implemented with fidelity is also of high priority (see Table 27). It would seem that the underlying focus is on mechanisms that support the change process/programme success and the ability to explain why the programme worked or did not work. This bottom-up approach mirrors a theory-driven approach to evaluation. I therefore decided to label this perspective as *theory-driven*.

Table 27

Factor 1 Crib Sheet: Brazil Cohort

Criteria	Item	Factor Score
Items with the highest rankings in Factor 1 array	Stakeholders agree on programme goals (QS4)	2.3
	Programme theory is plausible (QS 9)	2.0
	Programme theory is explicitly stated (QS8)	1.3
	Programme is implemented as intended (QS12)	1.2
Items ranked higher in Factor 1 array than any other factor array	Programme goals are clearly specified (QS1)	0.1
	Stakeholders agree on programme goals (QS4)	2.3
Items ranked lower in Factor 1 array than any other factor array	The manner in which the programme is delivered is clearly defined (QS10)	-0.5
	Stakeholders have authority to act on evaluation findings (QS14)	-1.3
	Stakeholders are transparent about the purpose of the evaluation (QS15)	-1.1
	Budget is adequate for the evaluation (QS16)	-0.5
	Type of evaluation required is feasible (QS18)	-0.8
Items with the lowest rankings in Factor 1 array	Stakeholders have authority to act on evaluation findings (QS14)	-1.3
	Stakeholders are transparent about the purpose of the evaluation (QS15)	-1.1

Nine respondents were most associated with the second factor. Evaluators who shared this perspective seemed to prioritise stakeholder transparency, authority and consensus, and fidelity of implementation (See Table 28). Issues pertaining to data collection (e.g., the quality and accessibility of evaluation data), evaluation design, and programme theory are assigned less importance. Given that the overall focus is on mechanisms that support the utilisation of findings, I decided to label this perspective as *utilisation-focused*.

Table 28

Factor 2 Crib Sheet: Brazil Cohort

Criteria	Item	Factor Score
Items with the highest rankings in Factor 2 array	Stakeholders have authority to act on evaluation findings (QS14)	2.1
	Stakeholders agree on programme goals (QS4)	1.6
	Programme is implemented as intended (QS12)	1.4
	Stakeholders are transparent about the purpose of the evaluation (QS15).	1.4
Items ranked higher in Factor 2 array than any other factor array	Programme outcomes are realistic (QS2)	0.9
	Stakeholders have authority to act on evaluation findings (QS14)	2.1
	Stakeholders are transparent about the purpose of the evaluation (QS15**)**	1.4
Items ranked lower in Factor 2 array than any other factor array	Programme outcomes are measurable (QS3)	-0.8
	Programme data are adequate (QS5)	-0.8
	Programme data are easily accessible (QS7)	-0.7
	Programme theory is explicitly stated (QS8)	-1.3
	Required evaluation methodology is feasible (QS19)	-0.8
Items with the lowest rankings in Factor 2 array	Programme theory is explicitly stated (QS8)	-1.3

The interpretation of Factor 3 and Factor 4 was attempted for exploratory purposes, keeping in mind that due to the relatively low proportion of variance accounted for by these two factors (13.0% and 10.4% respectively), it might be difficult to formulate concrete characterisations of these two perspectives based on the item configurations presented in Table 29 and Table 30.

Table 29

Factor 3 Crib Sheet: Brazil Cohort

Criteria	Items	Factor Scores
Items with the highest rankings in Factor 3 array	Programme theory is plausible (QS 9)	2.1
	Programme theory is explicitly stated (QS8)	1.5
	Stakeholders have authority to act on evaluation findings (QS14)	1.2
	Stakeholders are transparent about the purpose of the evaluation (QS15)	1.1
Items ranked higher in Factor 3 array than any other factor array	Programme data are reliable (QS6)	0.4
	Programme theory is explicitly stated (QS8)	1.5
	Programme theory is plausible (QS9)	2.1
	Timeframe is adequate to complete the evaluation (QS17)	0.2
Items ranked lower in Factor 3 array than any other factor array	Programme goals are clearly specified (QS1)	-1.3
	Programme outcomes are realistic (QS2)	-0.7
	Stakeholders agree on programme goals (QS4)	-1.6
	Target beneficiaries are clearly defined (QS11)	-1.3
Items with the lowest ranking in Factor 3 array	Stakeholders agree on program goals (QS4)	-1.6
	Programme goals are clearly specified (QS1)	-1.3
	Target beneficiaries are clearly defined (QS11)	-1.3

Table 30

Factor 4 Crib Sheet: Brazil Cohort

Criteria	Item	Factor Score
Items with the highest rankings in Factor 4 array	Programme is implemented as intended (QS12)	2.7
	The manner in which the program is delivered is clearly defined (QS10).	1.3
Items ranked higher in Factor 4 array than any other factor array	Programme data are adequate (QS5)	0.7
	Programme data are easily accessible (QS7)	0.8
	The manner in which the programme is delivered is clearly defined (QS10)	1.3
	Target beneficiaries are clearly defined (QS11)	0.3
	Programme is implemented as intended (QS12)	2.7
	Budget is adequate for the evaluation (QS16)	0.9
Items ranked lower in Factor 4 array than any other factor array	Programme data are reliable (QS6)	-1.2
	Programme theory is plausible (QS9)	-1.0
	Stakeholders are willing to collaborate with the evaluator (QS13)	-0.9
	Timeframe is adequate to complete the evaluation (QS17)	-0.8
Items with the lowest rankings in Factor 4 array	Stakeholders agree on programme goals (QS4)	-1.4
	Programme data are reliable (QS6)	-1.2
	Programme theory is plausible (QS9)	-1.0

If one focuses exclusively on the items with the highest and lowest rankings in the Factor 3 array, it would seem that evaluators who share this perspective prioritise an explicit and plausible change logic, and stakeholder authority and transparency over the clear definition of programme goals and target beneficiaries. If the entire item configuration is taken into account, the picture becomes more complex. It is clear that this perspective is not related to any specific notion of evaluation practice. For example, one would expect that a plausible change logic would include realistic

outcomes, and yet this particular item was ranked lower in Factor 3 array than any other factor arrays. In addition, two of the items with the highest factor scores in this array had a higher ranking in a different factor array. I decided to exclude Factor 3 from further analysis and discussion as I could not reconcile meaningfully the different items underling this perspective.

Interpretation of Factor 4 was less problematic (albeit not straightforward). The emphasis appears to be on the specification and proper implementation of the service delivery plan, availability and accessibility of data, and sufficient budget to conduct the evaluation. Data quality, plausibility of the change logic, and consensus on programme goals are of lower priority for the six evaluators who shared this perspective. The underlying focus appears to be on the minimum requirements to measure implementation fidelity (available and easily accessible data; and budget). I therefore decided to label this perspective as *implementation-focused*. Six respondents were most associated with this perspective.

One would expect that the overall profile of evaluators who use a theory-driven (cohort 1; *n* = 14), a utilisation-focused (cohort 2; *n* = 9) and an implementation-focused approach (cohort 3; *n* = 6) to assessing programme evaluability to be different. This was however not necessarily the case (see Table I17 in Appendix I). For example, the training of most evaluators in all three cohorts was limited to self-education and a short course certificate in evaluation (*n* = 10; *n* = 6; *n* = 4, respectively). What stands out is the difference in self-reported level of experience, with cohort 1 and cohort 3 having the lowest and highest level of experience respectively. Most evaluators (*n* = 9) in cohort 1 had a maximum of five years of experience in conducting evaluations, while the majority of evaluators in cohort 3 (*n* = 4) had at least 11 years of experience.

SA cohort: Evaluability perspectives and associated evaluator types.

Ten respondents were most associated with the first factor extracted for the SA cohort. Evaluators who shared this perspective seemed to prioritise: (a) an explicit and plausible theory, articulating realistic and measurable outcomes for a specific target population; and (b) data for measuring implementation fidelity (see Table 31).

Taken together, the underlying focus appears to be on opening the black box of evaluation, and the ability to explain why the programme worked or did not work. This perspective resonates with a theory-driven approach to evaluation, and was labelled as such.

Table 31

Factor 1 Crib Sheet: SA Cohort

Criteria	Item	Factor Score
Items with the highest rankings in Factor 1 array	Programme is implemented as intended (QS12)	1.1
	Programme outcomes are realistic (QS2)	1.1
	Programme theory is explicitly stated (QS8)	1.1
Items ranked higher in Factor 1 array than any other factor array	Programme goals are clearly specified (QS1)	0.9
	Programme outcomes are realistic (QS2)	1.1
	Programme outcomes are measurable (QS3)	0.3
	Programme data are reliable (QS6)	0.5
	Programme data are easily accessible (QS7)	0.6
	Programme theory is explicitly stated (QS8)	1.1
	Programme theory is plausible (QS9)	0.5
	Target beneficiaries are clearly defined (QS11)	0.5
Items ranked lower in Factor 1 array than any other factor array	Stakeholders have authority to act on evaluation findings (QS14)	-1.2
	Stakeholders are transparent about the purpose of the evaluation (QS15)	-1.6
	Budget is adequate for the evaluation (QS16)	-1.1
	Timeframe is adequate to complete the evaluation (QS17)	-1.8
Items with the lowest rankings in Factor 1 array	Timeframe is adequate to complete the evaluation (QS17)	-1.8
	Stakeholders are transparent about the purpose of the evaluation (QS15)	-1.6
	Stakeholders have authority to act on evaluation findings (QS14)	-1.2
	Budget is adequate for the evaluation (QS16)	-1.1

Six respondents were most associated with the second factor. Evaluators who shared this perspective seemed to prioritise stakeholder transparency, authority, consensus and collaboration (see Table 32). Issues pertaining to evaluation design, data collection, and evaluation timeframe are less of a priority. Given that the overall focus is on mechanisms that support the utilisation of findings, I labelled this perspective as *utilisation-focused*.

Table 32

Factor 2 Crib Sheet: SA Cohort

Criteria	Item	Factor Score
Items with the highest rankings in Factor 2 array	Stakeholders have authority to act on evaluation findings (QS14)	2.5
	The manner in which the program is delivered is clearly defined (QS10)	1.4
	Stakeholders are transparent about the purpose of the evaluation (QS15).	1.4
Items ranked higher in Factor 2 array than any other factor array	Stakeholders agree on programme goals (QS4)	0.6
	The manner in which the programme is delivered is clearly defined (QS10)	1.4
	Stakeholders are willing to collaborate with the evaluator (QS13)	0.7
	Stakeholders are transparent about the purpose of the evaluation (QS15)	1.4
Items ranked lower in Factor 2 array than any other factor array	Programme goals are clearly specified (QS1)	-0.4
	Programme outcomes are realistic (QS2)	-0.1
	Programme data are adequate (QS5)	-0.3
	Target beneficiaries are clearly defined (QS11)	-0.1
	Type of evaluation required is feasible (QS18)	-1.3
	Required evaluation methodology is feasible (QS19)	-1.7
Items with the lowest rankings in Factor 2 array	Required evaluation methodology is feasible (QS19)	-1.7
	Type of evaluation required (process, outcome or impact) is feasible (QS18)	-1.3
	Timeframe is adequate to complete the evaluation (QS17).	-1.0

When comparing the profiles of evaluators who used a theory-driven approach to assessing programme evaluability (cohort 1; *n* = 10) and those who used a utilisation-focused approach (cohort 2; *n* = 6), it is clear that most evaluators in cohort 1 were involved in a wider range of evaluation-related activities, compared to those in the second cohort (see Table I17 in Appendix I).

In terms of employment setting, most evaluators (*n* = 8) in cohort 1 were employed in either a non-profit/non-governmental organization or worked as an independent evaluation consultant, while most evaluators in cohort 2 (*n* = 4) were employed in a university setting. Cohort 1 also had a higher level of formal training in evaluation, with four out of ten evaluators having a Master's degree or PhD in programme evaluation. The training of most evaluators in cohort 2 (*n* = 5) was limited to self-education or a short course certificate in evaluation. The overall level of self-reported experience of both cohorts was however comparable, with most evaluators in cohort 1 (*n* = 6) and most evaluators in cohort 2 (*n* = 4) having at least six years of experience.

A summary of the results for research question 1 is presented below:

Nine factors/perspectives were retained for interpretation across the four evaluator cohorts of interest, four of which were unique perspectives. There was some overlap in terms of the essential evaluability criteria that defined these perspectives (see Table 33).

Table 33

Summary of Results for Research Question 1

Cohort	No. of factors retained	Criteria considered *Essential*	*n*	Perspective label	Profile of most evaluators sharing perspective
USA	2	Programme theory is explicitly stated (QS8) Programme is implemented as intended (QS12)	22	Theory-driven	Employed in public sector or as independent consultant (*n* = 13) High level of formal evaluation training (*n* = 13) Relatively high level of experience in conducting evaluation (at least 6 years of experience; *n* =16)
		Stakeholders are transparent about the purpose of the evaluation (QS15) Stakeholders have authority to act on evaluation findings (QS14) Programme outcomes are realistic (QS2)	22	Utilisation-focused	Employed in a university or non-profit context (*n* = 12) Training mostly limited to self-education or short-course in evaluation (*n* = 13) Relatively low level of experience in conducting evaluation (Between 1- 5 years of experience: *n* = 11)
UK	2	Programme is implemented as intended (QS12)	8	Theory-driven	No discernible pattern
		Stakeholders are transparent about the purpose of the evaluation(QS15) Stakeholders have authority to act on evaluation findings (Q14) Programme outcomes are realistic (QS2)	5	Theory and utilisation-focused	No discernible pattern

Table 33Table 23 cont.

Summary of Results for Research Question 1

Cohort	No. of factors retained	Criteria considered *Essential*	*N*	Perspective label	Profile of most evaluators sharing perspective
Brazil	3	Stakeholders agree on programme goals (QS4) Programme theory is plausible (QS9) Programme theory is explicitly stated (QS8) Programme is implemented as intended (QS12)	14	Theory-driven	Training limited to self-education or short-course in evaluation (*n* = 10) Relatively low level of experience in conducting evaluation (Between 0 to 5 years of experience; *n* = 9)
		Stakeholders have authority to act on evaluation findings (QS14) Stakeholders agree on programme goals (QS4) Programme is implemented as intended (QS12) Stakeholders are transparent about the purpose of the evaluation (QS15)	9	Utilisation-focused	Training limited to self-education or short-course in evaluation (*n* = 6) Relatively moderate level of experience in conducting evaluations
		Programme is implemented as intended (QS12) The manner in which the programme is delivered is clearly specified (QS10)	6	Implementation-focused	Training limited to short-course certificate evaluation (*n* = 4) Relatively higher level of experience in conducting evaluations (at least 11 years of experience: *n* = 4)

Table 33 cont.

Summary of Results for Research Question 1

Cohort	No. of factors retained	Criteria considered *Essential*	*N*	Perspective label	Profile of most evaluators sharing perspective
SA	2	Programme is implemented as intended (QS12) Programme outcomes are realistic (QS2) Programme theory is explicitly stated (QS8)	10	Theory-driven	Involved in a wider range of evaluation-related activities Employed by non-profit/non-governmental organization or working as an independent evaluation consultant (*n* = 8) High level of formal evaluation training (*n* = 5) At least 6 years of experience conducting evaluations (*n* = 8)
		Stakeholders have authority to act on evaluation findings (QS14) The manner in which the programme is delivered is clearly defined (QS10) Stakeholders are transparent about the purpose of the evaluation (QS15)	6	Utilisation focused	Employed in university setting (*n* = 4) Limited to self-education or short-course certificate in evaluation (*n* = 5) At least 6 years of experience conducting evaluations (*n* = 4)

A number of tentative conclusions can be drawn here. First, it is interesting to note that most USA and SA evaluators who used a theory-driven approach to assessing programme evaluability had similar training and experience profiles (i.e., relatively high levels of formal training in programme evaluation and experience in conducting evaluations). A reverse pattern was noted for Brazil evaluators who used a similar approach. Second, a utilisation-focused approach to assessing evaluability appears to be common amongst USA, SA and Brazil evaluators with a low level of formal training in programme evaluation. Third, an implementation-focused perspective appear to be unique to the Brazil evaluator cohort.

Are Evaluators' Prioritisation of Evaluability Criteria Consistent across Different Study Tasks?

Correspondence analysis (CA) was used to identify associations between a set of evaluability criteria and the different study tasks presented to the four cohorts of evaluators who participated in this study. The assumption was that prioritisation of particular evaluability criteria might vary depending on the nature of the study tasks. In the first three tasks, participants were required to respond to three different evaluation scenarios, each with a specific set of evaluability conditions. The fourth task was an acontextual exercise that required participants to sort a pre-determined set of evaluability criteria in order of importance.

Correspondence analysis was performed using four sets of data: the evaluability criterion that each respondent identified as *most important* in their separate assessments of Scenario 1, Scenario 2, and Scenario 3 (irrespective of whether the scenario was characterised as evaluable with ease or evaluable with difficulty); and the evaluability criterion that each respondent identified as *first on their priority list* in the Q Sort task.

The two-way contingency tables, which formed the basis of four separate correspondence analyses, are presented in Appendix J (see Tables J1-J4). The number of valid observations for the UK (n = 91) and SA (n = 141) cohorts were lower compared to the USA (n = 318) and Brazil (n = 230) cohorts. While it is easy to identify the evaluability criteria with the highest or lowest total frequency (i.e., rare

and common objects) by visually inspecting the data matrices, it is more difficult to analyse the overall structure of these matrices or dissect patterns of variations encoded in the data in this manner. The application of CA is therefore justified here as this technique would generate simultaneous visualisations of the row and columns categories, and allow me to isolate patterns of association within the data matrix at a category level.

Correspondence maps: USA cohort.

Three correspondence maps were generated for the USA cohort: one with all data points, one excluding low frequency points concentrated in only one study task, and one excluding rare objects identified by their outlying position in relation to other objects on the second map. Comparison of the three maps (see Figures J1-J3 in Appendix J) revealed that the relative positioning of the row and column categories changed substantially after excluding both QS10 (low frequency point) and QS12 (outlier) from the analysis. Results associated with the third correspondence map (see Figure 12) were therefore interpreted.

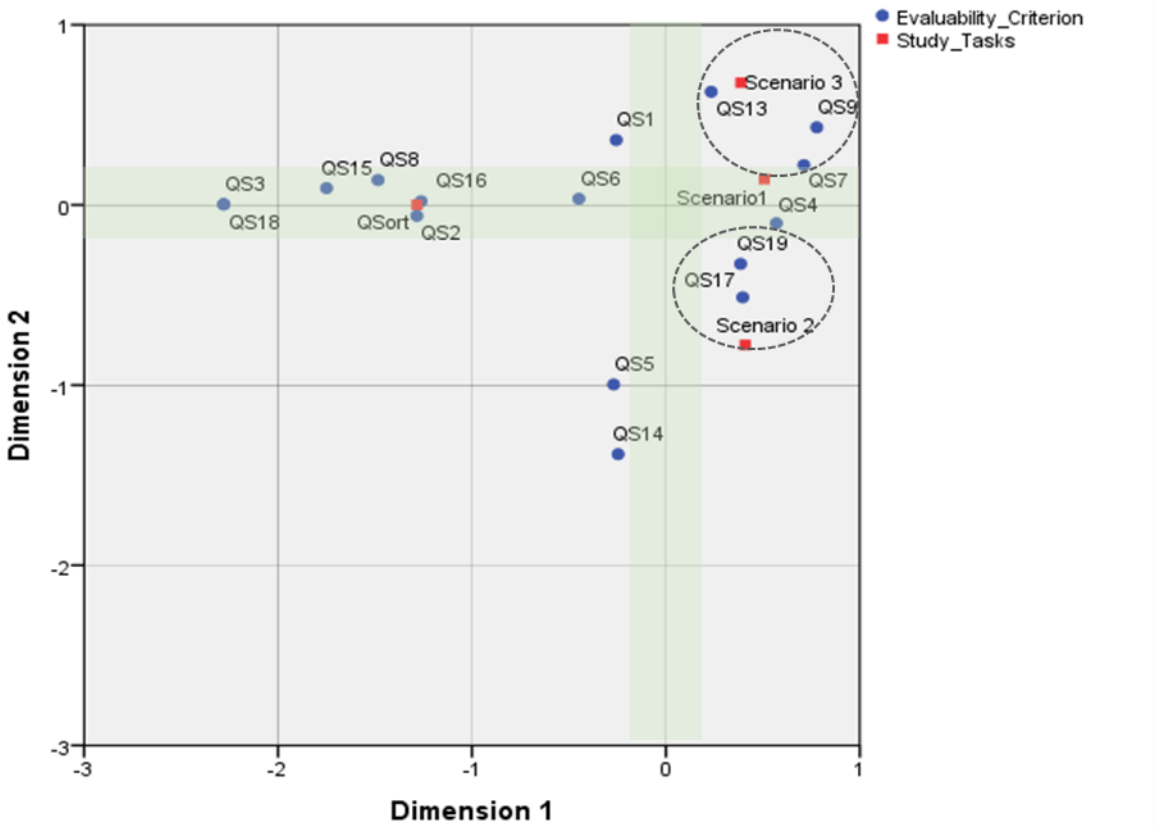


Figure 11. CA map excluding QS10 and QS12 (USA Cohort). The shaded area consists of points that cannot be meaningfully interpreted on one or both dimensions. QS11 does not appear on the map as it was not prioritised in any study task. Task-centroid lines, criterion-centroid lines, and criterion-task perpendiculars were omitted to reduce the complexity of the map.

There was a significant dependency between the row and column categories, χ^2 (48) = 140.4, $p < .05$ (see Table 34). This implies that the prioritisation of evaluability criteria and the type of study task were not independent of each other. It should be noted that three dimensions were extracted to explain the model, but only the first two dimensions were used to generate the correspondence map in Figure 11 (see chapter 4 for rationale). The first two dimensions accounted for 86.6% of the total variance explained by the model, with the first dimension explaining 71% of the variation.

The row and column coordinates for the two-dimensional solution are presented in Tables J4-J5 in Appendix J. The *Score in Dimension* column in each table displays

the score of each row and column on dimension 1 and dimension 2. These scores represent dimensional distances that were used to derive the two-dimensional biplot in Figure 12.

Table 34

Summary of CA Results: USA Cohort

					Proportion of Inertia		Confidence Singular Value	
Dimension	Singular Value	Inertia	χ^2	*p*	Accounted for	Cumulative	*SD*	Correlation
1	.563	.317			.710	.710	.045	0.18
2	.263	.069			.156	.866	.054	
3	.244	.060			.134	1.000		
Total		.446	140.422	.000	1.000	1.000		

I applied the guidelines described in Chapter 4 (under data analysis) to interpret Figure 12. I was particularly interested in row-to-column comparisons and column to column comparisons (i.e., how a particular study task compared to other study tasks).

Two distinct clusters, consisting of both row and column categories, can be identified in Figure 12. In the first cluster, QS19 (*required evaluation methodology is feasible*) and QS17 (*timeframe is adequate to complete the evaluation*) are close together on both dimensions, indicating that they have similar profiles across different study tasks. Based on the row-to-column distance interpretation method described in Chapter 4, one can conclude that these two evaluability criteria were prioritised more frequently in Scenario 2, compared to the other three study tasks. The profile of the QS17 was more prominent than that of QS19 in Scenario 2.

In the second cluster, QS13 (*stakeholders are willing to collaborate with the evaluator*), QS9 (*programme theory is plausible*), and QS7 (*Programme data are easily accessible*) were most frequently prioritised in Scenario 3 relative to the other study tasks. The profile of the evaluability criterion *stakeholders are willing to collaborate with the evaluator* (QS13) was more prominent in this particular study task.

Scenario 1, Scenario 2 and Scenario 3 are close together on dimension 1 (which explains most of the variability) but not on dimension 2, suggesting that these three study tasks were different in terms of the evaluability criteria that were prioritised across them. Scenario 1 and Scenario 3 are the closest on both dimensions suggesting that their profiles are the most similar. The Q Sort task is positioned the furthest from all other study tasks on dimension 1. This clear demarcation indicates that the Q Sort task had significantly different profiles from the other study tasks.

Correspondence maps: UK cohort.

The correspondence maps generated for comparison can be found in Appendix J (see Figure J4 and J5). The relative positioning of the row and column categories remained virtually the same after excluding low frequency points concentrated in

only one study task (QS2, QS9, and QS15). I decided to interpret the results associated with the second biplot (see Figure 13) for consistency.

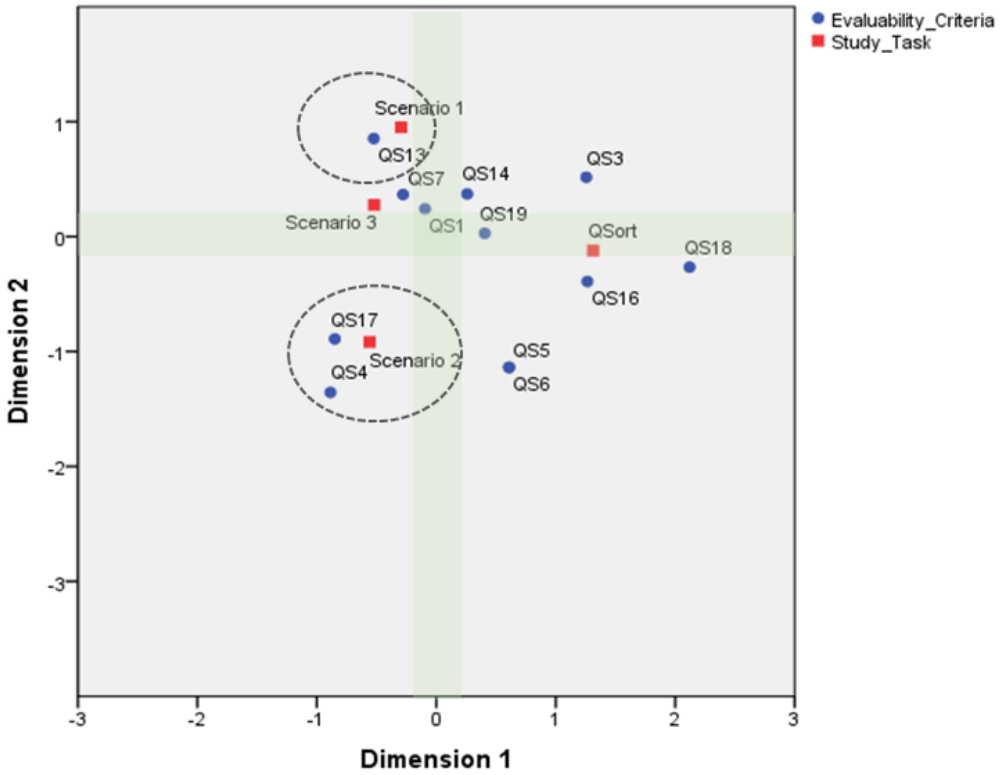


Figure 12. CA map excluding QS2, QS9 and QS15 (UK cohort). The shaded area consists of points that cannot be meaningfully interpreted on one or both dimensions. There are no discernible outliers. QS8, QS10, QS11, and QS12 do not appear on the map as they were not prioritised in any study task. Task-centroid lines, criterion-centroid lines, and criterion-task perpendiculars were omitted to reduce the complexity of the map.

While the dependency between row and column categories was not significant, $\chi^2(45) = 54.6$, $p > .05$, the total inertia value and the associated Phi coefficient (square root of total inertia value) was high, $ø = .8$ (see Table 35). This points to a strong association between row and column categories (Alberti 2013; Greenacre, 2007). I therefore deemed it reasonable to proceed with the interpretation of the biplot, keeping in mind the associated margin of error. The issue of statistical significance is assigned less importance in exploratory data analytic techniques such as CA, than in confirmatory or explanatory ones.

The two-dimensional solution accounted for 95.3% of the total inertia. Two distinct clusters can be identified in Figure 13. In the first cluster, QS13 (*stakeholders are willing to collaborate with the evaluator*) was prioritised more frequently in Scenario 1 compared to the other study tasks. In the second cluster, QS4 (*stakeholders agree on programme goals*) and QS17 (*timeframe is adequate to complete the evaluation*) were prioritised more frequently in Scenario 2 compared to the other study tasks. The profile of QS4 was however more prominent in Scenario 2.

As observed in the USA cohort, Scenario 1, Scenario 2 and Scenario 3 are close together on dimension 1 but not on dimension 2, with Scenario 1 and Scenario 3 being the closest on both dimensions. In addition, the Q Sort task is positioned the furthest from all other study tasks on dimension 1.

Table 35

Summary of CA Results: UK Cohort

					Proportion of Inertia		Confidence Singular Value	
Dimension	Singular Value	Inertia	χ^2	*p*	Accounted for	Cumulative	*SD*	Correlation
1	.619	.383			.617	.617	.071	.015
2	.457	.208			.336	.953	.088	
3	.171	.029			.047	1.000		
Total		.621	54.64	.154	1.000	1.000		

Correspondence maps: Brazil cohort.

The three correspondence maps generated for comparison can be found in Appendix J (see Figures J6-J8). Excluding low frequency points concentrated in only one study task (QS2, QS6, QS11, QS12, QS15, and QS18) substantially affected the positioning of the row and column categories. In contrast, excluding the two points with outlying positions (QS8, QS10) in the second correspondence map did not substantially affect the positioning of row and column categories. I therefore interpreted the results associated with the second map (see Figure 14).

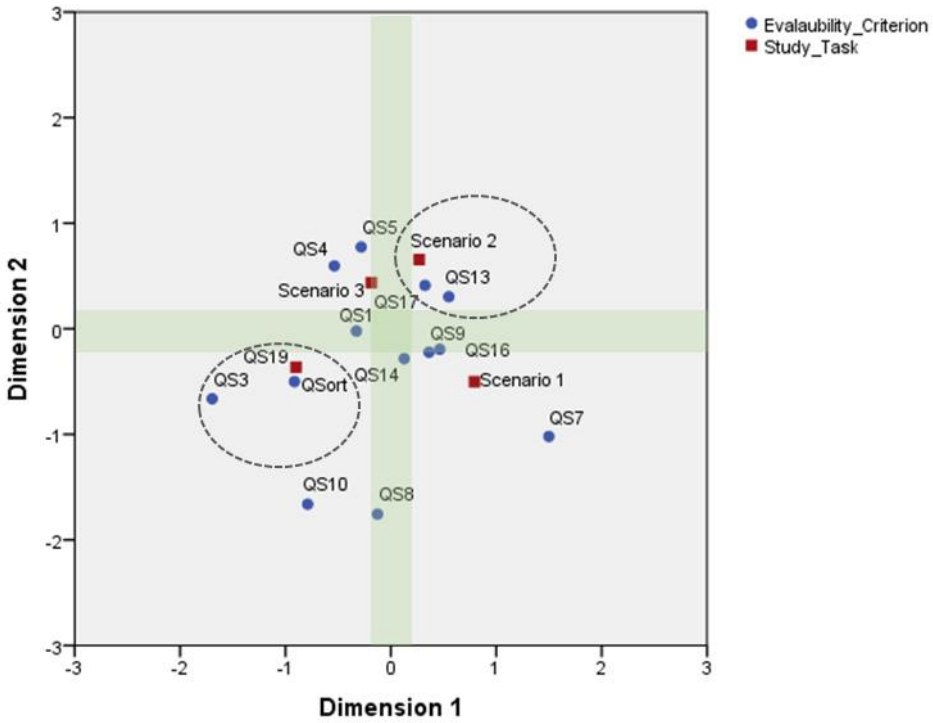


Figure 13. CA map excluding QS2, QS6, QS11, QS12, QS15 and QS18 (Brazil cohort). The shaded area consists of points that cannot be meaningfully interpreted on one or both dimensions. Task-centroid lines, criterion-centroid lines, and criterion-task perpendiculars were omitted to reduce the complexity of the map.

There was a significant dependency between the prioritisation of evaluability criteria and the type of study task, χ^2 (36) = 54.9, $p < .05$ (see Table 36). The model however accounted for only 26% of the variance in the correspondence table. The relatively low inertia can be attributed to the low Chi-square value and large number of observations. Three dimensions were extracted to explain the model, with the first

dimension explaining 69.5% and the second dimension explaining 23.3% of the total inertia. Cumulatively these two dimensions therefore explained 92.8% of the total variance explained by the model. This represents a significant proportion. The third dimension explained only 7.2% of the total variance, and was therefore deemed negligible.

Despite the low inertia, there is still a surprisingly clear pattern in the positions of certain study tasks and evaluability criteria. Scenario 2 lies in the upper right quadrant, Scenario 1 lies in the lower right quadrant, and the Q Sort task lies in the lower left quadrant. This clear demarcation indicates that these study tasks have significantly different profiles.

Two distinct clusters can be identified in Figure 14 upon the exclusion of points that could not be meaningfully interpreted. The different study tasks are well separated from one another: QS13 (*stakeholders are willing to collaborate with the evaluator*) and QS17 (*timeframe is adequate to complete the evaluation*) were prioritised more frequently in Scenario 2 compared to the other study tasks. In the second cluster QS19 (*required evaluation methodology is feasible*) and *QS3* (*programme outcomes are measurable*) are in close proximity to the Q Sort task. The profile of QS3 was more prominent in the Q Sort task than that of QS19.

Table 36

Summary of CA Results: Brazil Cohort

					Proportion of Inertia		Confidence Singular Value	
Dimension	Singular Value	Inertia	χ^2	*p*	Accounted for	Cumulative	*SD*	Correlation
1	.425	.181			.695	.695	.051	.242
2	.246	.061			.233	.928	.067	
3	.137	.019			.072	1.000		
Total		.260	54.882	.023	1.000	1.000		

Correspondence maps: SA cohort.

I interpreted the results associated with Figure 15 after deleting low frequency points (QS2, QS12, QS14, and QS18) and an outlier (QS3) from the analysis (see Figures J9-J10 in Appendix J).

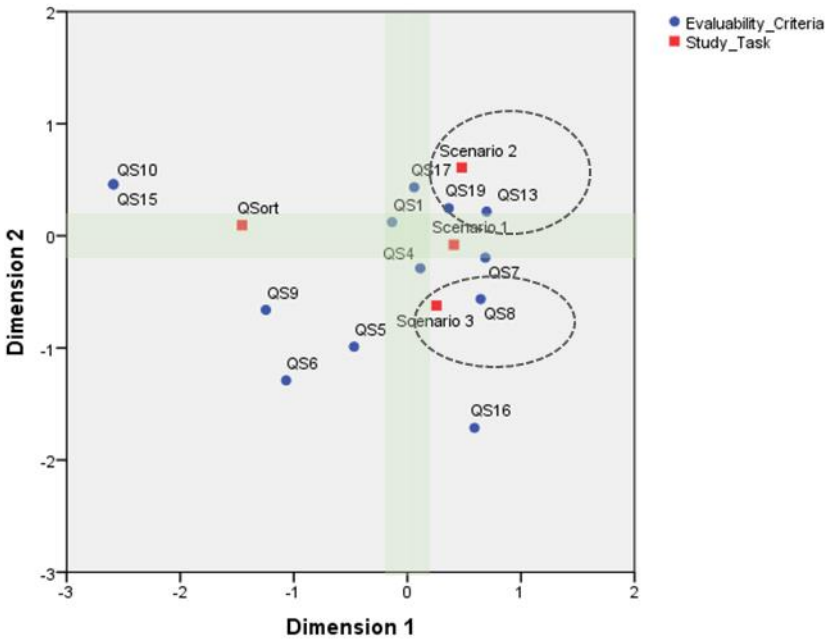


Figure 14. CA Map excluding QS2, QS12, QS14, QS18 and QS3 (SA Cohort). The shaded area consists of points that cannot be meaningfully interpreted on one or both dimensions. QS10, QS11, and QS15 do not appear on the map as they were not prioritised in any study task. Task-centroid lines, criterion-centroid lines, and criterion-task perpendiculars were omitted to reduce the complexity of the map.

While there was no significant dependency between the row and column categories, χ^2 (39) = 49.8, p > .05, the Phi coefficient was .6, thus pointing to a moderate association between the two categorical variables (see Table 37). The two-dimensional solution accounted for 93.6% of the total inertia. Two distinct clusters of correspondence can be identified in Figure 15 based on their relative proximity: QS19, QS13, and Scenario 2 in the upper right quadrant; and QS8 and Scenario 3 in the lower right quadrant. The profile of QS19 (*required evaluation methodology is feasible*) and QS13 (*stakeholders are willing to collaborate with the evaluator*) were

equally prominent in Scenario 2. QS8 (*Programme theory is explicitly stated*) was prioritised most frequently in Scenario 3 compared to other study tasks.

As observed in both the USA and UK cohorts, Scenario 1, Scenario 2, and Scenario 3 lie in close proximity on dimension but not on dimension 2, with the Q Sort task positioned the furthest from all other study tasks on dimension 1. In addition, Scenario 1 and Scenario 3 were the closest on both dimensions.

Table 37

Summary of CA Results: SA Cohort

					Proportion of Inertia		Confidence Singular Value	
Dimension	Singular Value	Inertia	χ^2	*p*	Accounted for	Cumulative	*SD*	Correlation
1	.562	.316			.826	.826	.064	-.162
2	.205	.042			.110	.936	.072	
3	.157	.025			.064	1.000		
Total		.383	49.787	.115				

A summary of the results for all evaluator cohorts is presented below. The following similarities were noted:

- Scenario 1, Scenario 2, and Scenario 3 lie in close proximity on the dimension that explains most of the variability (dimension 1), in the USA, UK, and SA correspondence maps. The Q Sort task is well separated from all study tasks in all three of these correspondence maps. This clear demarcation indicates that the Q Sort task had significantly different profiles from the other study tasks.
- Scenario 1 and Scenario 3 were consistently close on both dimension 1 and dimension 2 in the USA, UK, and SA correspondence maps. What differentiates Scenario 1 and Scenario 3 from Scenario 2 is the manipulation of unfavourable stakeholder characteristics in these two scenarios.
- The evaluability criteria *stakeholders are willing to collaborate with the evaluator* (QS13), *timeframe is adequate to complete the evaluation* (QS17), and *programme goals are clearly specified* (QS1) were the most prioritised criteria (in absolute terms) across all evaluator cohorts (See Tables J1-J4 in Appendix J). Stakeholders' willingness to collaborate with the evaluator was consistently prioritised more frequently than other stakeholder characteristics such as transparency and authority.
- The evaluability criteria *the manner in which the programme is delivered is clearly defined* (QS10) and *target beneficiaries are clearly defined* (QS11) were the least prioritised criteria (in absolute terms) across all evaluator cohorts. Both of these criteria fall under the evaluability category *programme design*.

The following differences were noted:

- Scenario 1, Scenario 2 and Scenario 3 are close together on dimension 1 but not on dimension 2 in the USA, UK, and SA correspondence maps, suggesting that these three study tasks were different in terms of the evaluability criteria that were prioritised across them.

- A different pattern emerged in the Brazil correspondence map: Scenario 1, Scenario 2, and Scenario 3 are widely dispersed on dimension 1. In addition, The Q Sort task is not as clearly demarcated from the other study tasks in this particular correspondence map, when compared to those derived for the USA, UK, and SA evaluator cohorts.
- The same criteria was not prioritised consistently in the same study tasks by all evaluator cohorts. For example, QS17 (*timeframe is adequate to complete the evaluation*) was prioritised more frequently in Scenario 2 compared to the other three study tasks by the USA evaluator cohort, while QS4 (*stakeholders agree on programme goals*) was prioritised more frequently in this particular scenario (relative to other study tasks) by the UK evaluator cohort.
- The relative prioritisation of evaluability criteria differed across study tasks, even within the same evaluator cohort. For example, the profile of QS17 (*timeframe is adequate to complete the evaluation*) was more prominent in Scenario 2, while that of QS13 (*stakeholders are willing to collaborate with the evaluator*) was more prominent in Scenario 3 for the USA evaluator cohort.

While there were some similarities in terms of which evaluability criteria were most frequently prioritised (in absolute terms) in the scenario tasks, there were vast enough differences in terms of relative prioritisation in the correspondence analysis to conclude that the evaluators' prioritisation patterns were not consistent across different study tasks.

Do Selected Evaluator Characteristics (practice context and experience) Predict Evaluators' Evaluability Assessments, Likelihood to Evaluate, and Prioritisation of Evaluability Criteria?

Multinomial Logistic Regression (MLR) was used to determine the extent to which level of experience and practice context predicted the evaluability criteria that evaluators prioritised in the three evaluation scenarios, evaluators' evaluability assessments and their decision to conduct an evaluation, given the specifics of each scenario. The results are organised under each dependent variable (DV) of interest. Before interpreting the results, I examined the standard errors associated with the *B*

coefficients in the parameter estimates tables derived for each analysis. None of the predictor variables had standard errors above two, thus ruling out the issue of multicollinearity.

Evaluator characteristics and prioritisation of evaluability criteria.

The proposed model did not significantly predict the DV better than the intercept-only model for any of the scenarios, χ^2 (8) = 10.9; χ^2 (8) = 12.3; χ^2 (8) = 8.2, $p > .05$ (see Table 38). There was no statistical evidence of the presence of a relationship between practice context and evaluator experience, and the DV (prioritisation of evaluability criteria). I therefore did not proceed with the MLR analysis.

Table 38

Model Fitting Information (DV: Prioritisation of Evaluability Criteria)

	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	χ^2	*df*	*p*
Scenario 1				
Intercept Only	64.053			
Final	53.130	10.923	8	.206
Scenario 2				
Intercept Only	72.054			
Final	59.790	12.264	8	.140
Scenario 3				
Intercept Only	71.041			
Final	62.836	8.206	8	.414

Evaluator characteristics and assessment of evaluability.

The proposed model, incorporating the two independent variables, significantly predicted the dependent variable better than the intercept-only model for Scenario 1, χ^2 (8) = 25.1, $p < .05$ (see Table 39). This was not the case for Scenario 2 and

Scenario 3. The proposed model however fitted the data well for all three scenarios (see Table K1 in Appendix K).

Table 39

Model Fitting Information (DV: Assessment of Evaluability)

	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	χ^2	*df*	*p*
Scenario 1				
Intercept Only	87.177			
Final	62.030	25.147	8	.001
Scenario 2				
Intercept Only	62.892			
Final	52.535	10.357	8	.241
Scenario 3				
Intercept Only	51.019			
Final	43.430	7.588	8	475

The model accounted for more variability in the assessment of Scenario 1 (between 10.5% and 11.9%) than Scenario 2 and 3 (see Table K2 in Appendix K for Pseudo R^2 statistics). Practice context was a significant predictor of evaluators' assessment of Scenario 1, χ^2 (4) = 14.0, $p < .05$, but not of Scenario 2 and Scenario 3 (see Table 40).

Table 40

Likelihood Ratio Tests (DV: Assessment of Evaluability)

	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	χ^2	*df*	*p*
Scenario 1				
Experience	71.438	9.408	4	.052
Practice Context	76.046	14.016	4	.007
Scenario 2				
Experience	54.641	2.106	4	.716
Practice Context	60.779	8.244	4	.083
Scenario 3				
Experience	45.184	1.753	4	.781
Practice Context	49.578	6.147	4	.188

The significance values associated with the Wald statistic were examined to determine whether or not practice context significantly differentiated between the categories of the dependent variable (see Table 41). Practice context (developed context in particular) significantly differentiated between the *Medium difficulty* and *High difficulty* category of the dependent variable (if the corrected *p* value of .025 is applied; see method chapter for rationale). Evaluators practising in developed countries were 5.5 times (1/.183) more likely to characterise Scenario 1 as *evaluable with high difficulty* than as *evaluable with medium difficulty*, $p < .025$, 95% CI [.51, .66].

Table 41

Parameter Estimates for Scenario 1 (DV: Evaluability Assessment)

Scenario 1 Evaluability level		*B*	Std. Error	Wald	*Df*	*P*	Exp (*B*)	95% Confidence Interval for Exp(*B*)	
								Lower Bound	Upper Bound
Low	Intercept	-1.591	.847	3.530	1	.060			
	Low Experience	.845	.366	5.337	1	.021	2.32	1.137	4.771
	Medium Experience	.718	.440	2.662	1	.103	2.050	.865	4.854
	High Experience	0[a]	.	.	0	.	.	.	.
	Developing Context	1.069	.843	1.611	1	.204	2.913	.559	15.190
	Developed Context	.439	.853	.265	1	.607	1.552	.292	8.257
	Both	0[a]	.	.	0	.	.	.	.
Medium	Intercept	-.138	.608	.052	1	.820			
	Low Experience	.343	.438	.614	1	.433	1.409	.597	3.326
	Medium Experience	.998	.483	4.267	1	.039	2.712	1.052	6.989
	High Experience	0[a]	.	.	0	.	.	.	.
	Developing Context	-.752	.617	1.490	1	.222	.471	.141	1.578
	Developed Context	-1.699	.653	6.761	1	.009	.183	.051	.658
	Both	0[a]	.	.	0	.	.	.	.

Note: The first set of coefficients/logits represent the comparison of the Low evaluability category to the reference category (High), and the second set of coefficients represent the comparison of the Medium evaluability category to the reference category (High).

[a]This parameter is set to zero because it is redundant.

Overall, the model accurately predicted 52%, 70.2%, and 79.7% of the cases for Scenario 1, Scenario 2, and Scenario 3 respectively (see Tables K3-K5 in Appendix K). The proportional by chance accuracy criteria was 45% for scenario 1, 67.5% for scenario 2, and 81.3% for scenario 3 (See Tables K6-K8 and relevant calculations in Appendix K). The overall criterion of classification accuracy was therefore satisfied for Scenario 1 and Scenario 2, suggesting the model was useful in predicting the cases for these two scenarios (i.e., the explanatory variables contributed to the explanation of the dependent variable in Scenario 1 and Scenario 2).This was however not the case for Scenario 3.

Evaluator characteristics and likelihood of conducting evaluation.

The proposed model significantly predicted the dependent variable better than the intercept-only model for Scenario 1, χ^2 (8) = 31.3, $p < .05$ (see Table 42). There was statistical evidence of the presence of a relationship between the predictor variables (practice context and evaluator experience) and the likelihood of conducting an evaluation for Scenario 1. This was not the case for Scenario 2 and 3. The proposed model fitted the data well (see Table K9 in Appendix K).

Table 42

Model Fitting Information (DV: Likelihood of Conducting Evaluation)

	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	χ^2	*df*	*p*
Scenario 1				
Intercept Only	90.048			
Final	58.768	31.279	8	.000
Scenario 2				
Intercept Only	65.479			
Final	56.432	9.047	8	.338
Scenario 3				
Intercept Only	52.787			
Final	52.339	.448	8	1.000

The proportion of variance explained by the model was negligible for Scenario 2 and 3. On the other hand, the model explained between 12.8% and 14.5% of the variability in the DV for Scenario 1 (see Table K10 in Appendix K).

Experience level and practice context were statistically significant predictors of evaluator's likelihood to evaluate Scenario 1, χ^2 (4) = 13.7; χ^2(4) = 14.7, $p < .05$. Only practice context was a significant predictor of evaluator's likelihood to evaluate Scenario 2, χ^2 (4) = 13.7, $p < .05$. None of the predictor variables approached statistical significance for Scenario 3 (see Table 43).

Table 43

Likelihood Ratio Tests (DV: Likelihood of Conducting Evaluation)

	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	χ^2	*df*	*p*
Scenario 1				
Experience	72.460	13.692	4	.008
Practice Context	73.486	14.718	4	.005
Scenario 2				
Experience	57.507	1.075	4	.898
Practice Context	70.100	13.688	4	.008
Scenario 3				
Experience	57.797	5.457	4	.243
Practice Context	58.068	5.729	4	.220

While both experience level and practice context were statistically significant predictors of evaluator's likelihood to evaluate Scenario 1, only experience level (low experience level in particular) was significant in distinguishing the *Low likelihood* and *High likelihood* category of the dependent variable, even if a less stringent significance level of .05 is applied (see Table 44). Evaluators with low experience level (≤ 1 year to 5 years) were 2.5 times (1/.395) more likely than unlikely to evaluate Scenario 1, $p < .025$, 95% CI [.19, .80].

Practice context did not significantly distinguish between the different categories of the DV for Scenario 2 (see Table K11 in Appendix K).

Table 44

Parameter Estimates for Scenario 1 (DV: Likelihood of Conducting Evaluation)

Likelihood of evaluating program (Scenario 1)[a]		*B*	Std. Error	Wald	*df*	*p*	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Low likelihood	Intercept	.349	.628	.308	1	.579			
	Low Experience	-.929	.366	6.453	1	.011	.395	.193	.809
	Medium Experience	-.419	.446	.880	1	.348	.658	.274	1.577
	High Experience	0[a]	.	.	0	.	.	.	.
	Developing Context	-.202	.644	.098	1	.754	.817	.231	2.885
	Developed Context	.830	.652	1.622	1	.203	2.294	.639	8.232
	Both	0[a]	.	.	0	.	.	.	.
Moderate Likelihood	Intercept	-.561	.695	.651	1	.420			
	Low Experience	.223	.453	.242	1	.623	1.250	.514	3.036
	Medium Experience	.917	.515	3.168	1	.075	2.501	.911	6.865
	High Experience	0[a]	.	.	0	.	.	.	.
	Developing Context	-.111	.668	.028	1	.868	.895	.242	3.314
	Developed Context	-.331	.716	.214	1	.644	.718	.176	2.923
	Both	0[a]	.	.	0	.	.	.	.

Note: The first set of coefficients/logits represent the comparison of the Low likelihood category to the reference category (High), and the second set of coefficients represent the comparison of the Moderate likelihood category to the reference category (High).

[a]This parameter is set to zero because it is redundant.

Overall, the model accurately predicted 52.6%, 60%, and 64.1% of the cases for Scenario 1, Scenario 2, and Scenario 3 respectively (See Tables K12-K14 in Appendix K). The proportional by chance criteria was 43.8% for scenario 1, 55% for scenario 2, and 61.3% for scenario 3 (see Tables K15-K17 in Appendix K). The classification accuracy rate for all three scenarios were above the proportional by chance criteria, confirming the usefulness of the regression model.

In summary, the results of the MLR suggest that practice context and level of experience did not predict the type of evaluability criteria prioritised in any of the scenarios. Practice context was however a significant predictor of evaluators' overall assessment of Scenario 1, with evaluators practising in developed countries being 5.5 times more likely to characterise Scenario 1 as *evaluable with high difficulty* than medium *difficulty*. Evaluators with low experience level were 2.5 times more likely than unlikely to evaluate Scenario 1. It is interesting to note that predictive relationships were only identified in the context of Scenario 1, suggesting that the nature of the scenario might have influenced the response pattern of evaluators.

CHAPTER SIX

DISCUSSION AND CONCLUSIONS

This exploratory study adds to the recent surge in research on evaluation practice by identifying systematically: (a) evaluators' perspectives on programme evaluability, (b) the criteria that evaluators use to assess programme evaluability in response to different evaluation scenarios, and (c) background characteristics that predict programme evaluability decisions. The evaluation practice of four different cohorts of evaluators were empirically investigated and contrasted in this study. Participating evaluators either practised in a developed country with a mature evaluation culture (USA or UK), or in a developing country with a weak evaluation culture (Brazil or SA). This study represents to my knowledge the first theoretically and empirically grounded investigation of how evaluators from four different countries assess programme evaluability.

This chapter synthesises and explains the results of the study, and their practical and theoretical implications/contributions. Directions for future research are also discussed.

Evaluability Perspectives and Evaluator Types

The primary research question was: Do evaluators share a common perspective towards evaluability? If not, what perspectives can be empirically identified and what evaluator types are most associated with these perspectives? This question can be addressed by interpreting the results presented in Table 33.

The results of this study demonstrate that participating evaluators did not share a unified perspective towards evaluability. Four empirically distinct perspectives emerged from the data, suggesting that evaluators may approach evaluability assessments differently. These perspectives were labelled as *theory-driven*, *utilisation focused, implementation-focused*, and *theory and utilisation-focused*. The first perspective (*theory-driven*) was shared by all four evaluator cohorts, and can thus be considered as the most dominant perspective (52 evaluators defined this

particular perspective). The second perspective (*utilisation-focused*) was shared by at least one group of evaluators from the USA (*n* = 22), Brazil (*n* = 9), and SA (*n* = 5). Although each of these perspectives were characterised by slightly different item configurations, their underlying thrust was the same across the relevant evaluator cohorts. The *implementation-focused* perspective and the combined *theory and utilisation-focused* perspective were unique to the Brazil and UK cohorts, respectively.

Four distinct evaluator types were identified in this study: (a) evaluators with a high level of formal training and experience in evaluation, (b) evaluators with a low level of formal training and experience in evaluation, (c) evaluators with a low level of formal training but high level of experience in evaluation, and (d) evaluators with a low level of formal training but varied levels of experience in evaluation (i.e., a mix of low, moderate, and high levels of experience). Most USA and SA evaluators who shared a *theory-driven* perspective could be classified under the first evaluator type. The second evaluator type was more predominant amongst Brazil evaluators who shared a *theory-driven* perspective. A *utilisation-focused* perspective emerged primarily amongst USA, SA and Brazil evaluators who had a low level of formal training but varied levels of experience (fourth evaluator type), while an *implementation-focused* perspective was common amongst Brazil evaluators with a low level of formal training but high level of experience (third evaluator type).

Several general conclusions can be drawn from the results presented above. First, the mental models/perspectives of evaluators within each evaluator cohort were quite different, with evaluators from Brazil having the most divergent perspectives on evaluability. Mental models are representations of reality that are shaped by deeply ingrained assumptions or generalisations. These organised knowledge structures influence our understanding of a particular phenomenon and the associated decisions that we make (Mathieu, et al., 2000). The terms *mental models*, *mindsets*, and *perspectives* are used interchangeably in the literature, and their distinctions are not clear (Duffy, 2009). Argyris and Schon (1978, as cited in Duffy, 2009) refer to these closely intertwined concepts as *espoused theories of action*.

Second, the finding that certain evaluability perspectives were shared by all evaluator cohorts suggests that the views of a select group of evaluators were compatible, even if they did not practise in the same context. Third, results of this study suggest that perspectives on evaluability are shaped, in part, by level of experience and formal training, and not necessarily practice context. Each of these conclusions and associated implications are discussed next.

Divergent/multiple evaluability perspectives *within* evaluator cohorts: Reasons, implications, and solutions.

In this section, I address four questions that stem logically from the first finding of this study:

(a) Why were divergent/multiple evaluability perspectives identified within each evaluator cohort?
(b) What are the implications of having multiple/divergent evaluability perspectives on our discipline and practice?
(c) How can multiple/divergent evaluability perspectives be reconciled?
(d) Should we have a unified perspective on evaluability?

Why were divergent/multiple evaluability perspectives identified within each evaluator cohort?

The first question can be addressed by examining the evolution of programme evaluation. When the field emerged in the 1960s, it was dominated by a single overarching paradigm, grounded in Campbell's (1969) vision of the *Experimenting Society*, namely the application of rigorous quantitative methods to determine programme effectiveness (Palumbo & Nachmias, 1983). As the field matured, there was a surge in competing paradigms addressing different areas of evaluation practice. The evolving evaluation landscape was at some point “marked by vitality and disorder. The scale, ubiquity, and diversity of evaluation activities [made] comprehension difficult, even for those operating within the field." (House, 1980, p. 11, as cited in Palumbo & Nachimas, 1983). Four overarching paradigms, with clear philosophical orientations, can now be identified in the contemporary evaluation

literature (Mertens & Wilson, 2012): post-positivist (methods-oriented), pragmatic (use-oriented), constructivist (value-engaged), and transformative (social justice oriented). Evaluation is therefore a multi-paradigm discipline, with different co-existing schools of thought/intellection traditions. This is in fact the norm within other well-established disciplines like psychology, sociology, and organizational sciences (Cooper, 2014; Pfeffer, 1993).

The existence of divergent perspectives towards programme evaluability can be attributed to the nature of our discipline, and the large number of practitioners in the field of programme evaluation. When only a small number of practitioners work within a discipline or sub-discipline, the entire community tends to subscribe to one particular school of thought to ensure that progress is made (Cooper, 2014).

One might argue that the existence of competing perspectives towards evaluability is indicative that the evaluation community does not have a clear and collective understanding of this particular construct. I suggest that this position is acceptable for two additional reasons: (a) there is a lack of consensus on the working definition of evaluability in the literature, and (b) the extant grey literature is fraught with debates over the fundamentals of evaluability (for e.g., some evaluators question the need to assess evaluability given that any programme can be subjected to some form of evaluation; others question whether or not evaluability can be measured).

What are the implications of having multiple/divergent evaluability perspectives on our discipline and practice?

The second question can be addressed by examining the pragmatic challenges of having an evaluation community characterised by multiple/divergent perspectives towards programme evaluability. One can reasonably argue that the practice of evaluators who share a common perspective is informed by a common set of underlying values. They are most likely to agree on a number fundamental issues such as, what counts as good practice, what questions are worth investigating, and what methods are to be used. Divergent perspectives towards evaluability could therefore have a number of negative implications on collaborative work. For example, more time might be needed to resolve fundamental differences in

approaches or concept definition, interdependent activity might become more difficult to coordinate, and efficiency might be comprised by greater task uncertainty (Pfeffer, 1993). Many evaluations are collaborative ventures, conducted by multi-disciplinary teams. While this is an untested notion, it is conceivable that evaluators with divergent perspectives towards evaluability might find it difficult to work collaboratively on an evaluability assessment. For instance, evaluators with a *theory-driven* perspective might consider the assessment of stakeholders' level of instrumental authority and potential collaboration as a waste of valuable resources, while their counterparts who share a *utilisation-focused* perspective might lobby for such an assessment. The situation becomes even more challenging if an evaluation team consists of evaluators with fragmented perspectives on evaluability. The finding that 62% of participating evaluators (n = 260) in this study had fragmented perspectives lends credence to this possibility.

An evaluation community characterised by a lack of consensus on what constitutes an evaluable programme can stagnate in terms of skills and knowledge development (Brunner, 2006), especially if there are minimal attempts to integrate or resolve fundamentals differences across evaluators (Cooper, 2014; Pfeffer, 1994; Shadish & Epstein, 1987). At present, there is limited dialogue amongst evaluators with divergent perspectives on evaluability, and a thin empirical base to assess the merits of each perspective.

How can multiple/divergent evaluability perspectives be reconciled?

The third question can be addressed by examining the strategies that contenders of other multi-paradigm disciplines (e.g., the Psychological sciences) use to communicate and coordinate their actions. Here I draw on Cooper's (2014) ideas to discuss how dialogue and collaboration can be facilitated amongst evaluators with divergent perspectives on evaluability. The first applicable strategy is to fix the meaning of the term *evaluability* as the concept is articulated in ambiguous and inconsistent terms in the literature (Trevisan, 2007). The meaning of the term can be fixed by creating and validating prototypical examples of *unevaluable* programmes or programmes that are evaluable with difficulty. Such prototypes can be similar to the scenarios designed for the purpose of this study. According to Cooper (2014, p. 99),

“when the meaning of a term is fixed by pointing at an example of a kind, the fact that different [practitioners] may have different beliefs about things of that kind is irrelevant. Regardless of their different beliefs, all speakers talk about the same thing”.

The second applicable strategy for facilitating collaboration between evaluators with divergent evaluability paradigms is to interact with evaluation stakeholders as a team throughout the evaluability assessment process. This approach will ensure that all evaluators have direct access to the same contextual information, form a common frame of reference, and make a joint decision about evaluability (task delegation or role differentiation might be counter-productive in this context; see Klimoski & Mohammed, 1994; Levesque, Wilson & Wholey, 2001).

The third applicable strategy is to use the evaluability criteria imposed by external regulating bodies to encourage evaluators to set aside their conflicting perspectives for the purpose of collaborative work (in the Psychological Sciences the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders is used for this purpose). The issue here is that different funding or development agencies have different evaluability checklists and scoring protocols. How do we decide which one to use? Even if we do select the most comprehensive checklist, which evaluability dimension/criterion should be assigned the highest weighting? Most evaluability checklists do not have an explicit weighting system (Davies, 2013). One exception is the ILO evaluability assessment tool, in which the raw score on selected evaluability dimensions is multiplied by a specific weight/ratio, representing the relative priorities assigned to each dimension. The weight assigned to each dimension (objectives: 15%; indicators: 25%; baselines: 20%; milestones; 10%; risks and assumptions: 15%; and monitoring and evaluation: 5%) were determined based on “the expertise, experiences, and best practices” of the ILO’s Evaluation Office (International Labour Organization, 2012, p. 2).

There is a clear disparity between the evaluability criteria given the highest priority in this study (*stakeholders are willing to collaborate with the evaluator*, *timeframe is adequate to complete the evaluation*, and *programme goals are clearly specified)* and those assigned the highest cumulative weighting in the ILO evaluability

assessment tool (and vice versa). This disparity is to be expected as the ILO is narrowly focused on complying with its Results Based Management approach, hence its strong emphasis on indicators and baselines. The evaluability dimensions/criteria derived and validated in this study are, on the other hand, not tied to a specific approach.

Now that we know which evaluability criteria have been assigned the most/least importance by evaluators in this study, should we motivate for weights to be assigned accordingly (to avoid the mistaken assumption that all evaluability criteria are equally important)? For example, should the evaluability criteria *stakeholders are willing to collaborate with the evaluator* and *timeframe is adequate to* be assigned the highest weightings? Not necessarily. Davies and Payne (2015) proposed a more flexible alternative: instead of building pre-defined weights into selected evaluability instruments, the onus is on evaluators conducting an evaluability assessment to determine the weight that each dimension/criterion should carry. This would ensure that the assigned weights are sensitive to the evaluation context, the specific type of evaluation to be conducted (e.g., formative vs. summative) and the evaluation approach to be used. For example, if the preferred approach is a goal-free evaluation (GFE), it would make sense to assign relatively low weights to outcome definition and plausibility of programme theory (Davies, 2016; Youker, 2013). In the same vein, if the desired evaluation type is an impact evaluation, it would make sense to assign a relatively high weighting to the evaluability criterion *required evaluation methodology is feasible*, as stringent methodological requirements are required for this type of evaluation

Should we have a unified perspective on evaluability?

The fourth question is a contentious one. As previously argued, multiple/divergent perspectives can be adaptive as long as there is some agreement on the fundamentals (Cooper, 2014; Pfeffer, 1994). Very few professions are characterised by practitioners who rigidly adhere to one distinct perspective on an issue, and according to Shadish and Epstein (1987), there is no obvious motivation for programme evaluators to be any different. In addition, this state of affairs is sustainable as long as there is a sufficient number of practitioners accumulating

empirical knowledge on each perspective. While we do need to have a unified perspective on evaluability (given the range of acceptable evaluation approaches, purposes, and methods), we do need to deliberate on the intricacies of each perspective. Here I am not suggesting that we decide which evaluability perspective is better; the evaluability perspectives that emerged in this study are fundamentally ideological, and cannot be tested in a replicable manner (Coryn, Noakes, Westine, & Schroter, 2011). I am simply suggesting that we accumulate empirical knowledge on the practice of evaluators with different evaluability perspectives (for e.g., one can investigate the inherent challenges associated with each perspective; one can also examine how evaluators overcome these challenges when conducting evaluability assessments). It is reasonable to assume that using a particular approach to evaluability represents a rational response to a practical situation faced by evaluators. A detailed explication of that rationale was beyond the scope of this study–a gap that future research can address. As Shadish, Cook, and Leviton (1991, p. 31) remarked, "the ideal (never achievable) evaluation theory would describe and justify why certain evaluation practices lead to particular kinds of results across situations that evaluators confront". As researchers, we need to work towards achieving this ideal.

Shared evaluability perspectives *across* evaluator cohorts: Characterisation and implications.

This section addresses the second finding of this study: evaluability perspectives shared by all/most evaluator cohorts. The discussion is framed around the following questions:

(a) What principles underlie the main evaluability perspectives identified in this study?
(b) What factors accounted for the emergence of these perspectives?
(c) What are the implications of having shared evaluability perspectives across evaluator cohorts?

What principles underlie the main evaluability perspectives identified in this study?

The dominant perspective that emerged across all cohorts of interests was labelled as *theory-driven*. The hallmark of this perspective is its emphasis on unpacking "programmatic black boxes and [explaining] how and why programs work (or fail to work) in different contexts and for different program stakeholders" (Astbury & Leeuw, 2010, p. 364). This perspective aligns with Epstein and Klerman's (2012) proposed logic model approach to evaluability. This approach requires the explicit specification of a programme's theory of change in the form of a "falsifiable logic model" (Epstein & Klerman, 2012, p. 380). A falsifiable logic model (FLM) is an extension of the conventional logic model in that it contains extensive process-related detail, and quantitative benchmarks for programme operations and intermediate outcomes. An example of intermediate benchmarks for a training programme can be expressed as follows: classes of specified size are recruited (enrolment standards are specified), instructors teach with fidelity to the curriculum (a precise operational definition of fidelity is provided), students attend most of the classes (a quantitative standard is specified), and students master material (improvement is measured in terms of *X* percent gain on a post-test). The underlying assumption of this approach to evaluability is that programmes that do not satisfy the requirements of their own FLM are not ready for rigorous impact evaluations (RIE). Many programmes fall short of this expectation. Common forms of logic model failures include: (a) failure to secure required inputs, (b) low programme enrolment/demand for the programme, (c) low programme engagement/completion rates, (d) low fidelity, and (e) minimal progress on pre/post measures. These logic model failures disrupt service delivery, dilute programme impact, and complicate the design of a RIE.

Epstein and Klerman (2012) argue that while the FLM approach to evaluability might lengthen the evaluation timeline, it is an inexpensive approach, which relies extensively on existing programme operating/process records and post-programme measurements.

Epstein and Klerman's (2012) notion of an augmented logic model is at the core of recent conceptualisations of theory-driven evaluations. This form of evaluation is

driven by “contextualised, comprehensive, [and] ecological program theory models” (Coryn et al., 2011, p. 202) in an attempt to address the “black box problem” (Astbury & Leeuw, 2010, p. 364). Regardless of the way the augmented logic model is explicated and depicted during the conceptual phase of theory-driven evaluations, it essentially represents a “plausible and sensible model of how a program is supposed to work” (Bickman, 1987, p. 5). The systematic testing (plausibility check) and refinement of this model occurs during the empirical phase of theory-driven evaluations (Rogers, Petrosino, Huebner, & Hacsi, 2000). According to Coryn et al. (2011), there are five salient principles (and associated sub-principles) that demarcate theory-driven evaluations from other forms of evaluations, such as empowerment evaluation and goal-free evaluation. These principles were vetted by leading scholars as part of their study and include: (a) programme theory formulation; (b) theory-guided question formulation/prioritisation; (c) theory-guided evaluation design, planning, and execution; (d) theory-guided construct measurement; and (e) causal description and explanation, with an emphasis on the latter. These principles can be reduced to Donaldson’s (2007, p. 10) simplified three-step approach to “programme theory-driven evaluation science”: (a) developing programme impact theory, (b) formulating and prioritising evaluation question, and (c) answering evaluation questions.

Epstein and Klerman’s (2012) notion of an augmented logic model is also at the core of Pawson and Tilley’s (2009) realist approach to evaluation. In a realistic evaluation, it is not sufficient to causally link programmes to outcomes; identifying the underlying “mechanisms [that] are fired in [particular] contexts to produce outcomes” and “the pre-existing structures that enable or disable the intended mechanism of change” (Pawson & Tilley, 2009, p. 85; p.71) are key. Realistic evaluations therefore attempt to unpack the context-mechanism-outcome configuration (CMOC) by developing and testing CMOC theories. Realistic evaluation is therefore a type of theory-driven evaluation. The only distinction is that “the constituents of the theories are specified in realist terms” (Tilley, 2000, p. 7).

The second evaluability perspective that emerged consistently across three evaluator cohorts was labelled as *utilisation-focused*. The hallmark of this paradigm was stakeholder authority, transparency, and consensus. The contentious label

utilisation-focused was used to characterise this perspective as it consisted of empirically supported factors that promote evaluation use (some authors prefer the term *influence* over *utilisation* or *use;* see Kirkhart, 2000). In their review of 41 empirical studies conducted over a 25 year period, Johnson et al. (2009) found that stakeholder involvement in the evaluation process, and stakeholder-evaluator interaction and communication are key to maximising evaluation use. This finding aligns with one of main premise underlying Patton's (2008) utilisation-focused approach: evaluation is more likely to be used if primary intended users are involved in a meaningful manner in the evaluation process, feel ownership of the process, and have a stake in the findings. While stakeholder dynamics are not the only factors that have been linked to evaluation use (see Cousins and Leithwood, 1986), over half of the studies (23 out of 41) included in Johnson's et al. (2009) review investigated this particular factor, and the bulk of these studies supported its relationship with other use factors.

The centrality of use in our evaluation practice is well recognised and has led to the development of participatory, stakeholder-based, and collaborative approaches to evaluation (Ayers, 1987; Cousins & Earl, 1992; Kirkhart, 2000; O'Sullivan, 2012). The *utilisation-focused* evaluability perspective that emerged in this study aligns with these approaches. The stakeholder-based approach was introduced to reconcile "varied political perspectives through interactive processes and [incorporate] multiple viewpoints into the design and conduct of the evaluation" (Cousins & Earl, 1992, p. 399). The participatory evaluation approach is an extension of the stakeholder-based approach and can be distinguished from the latter by the extent and scope of involvement required from primary intended users. Proponents of this approach are more concerned with enhancing evaluation use than neutralising stakeholder differences. The evaluation literature is inundated by variants of both of these approaches (e.g., practical and transformative strands of participatory evaluation; see Brisolara, 1998; Cousins & Whitmore, 1998), all built on the underlying principle of extensive stakeholder involvement (Brandon & Fukunaga, 2014; Daigneault, Jacob, & Tremblay, 2012). According to King (1998, p. 58), the "profusion of terms [to designate methods characterised by stakeholder involvement, such as responsive evaluation, empowerment evaluation, and deliberative-democratic

evaluation] is surely an indication that participatory approaches to program evaluation are coming of age".

What factors accounted for the emergence of these perspectives?

Now that we have characterised the dominant evaluability perspectives (*theory-driven* and *utilisation-focused*) that emerged in this study and connected them to existing notions of evaluation, we can isolate factors that might have accounted for their consistent emergence across evaluator cohorts. First, it is conceivable that these perspectives emerged because notions of "unpacking the black box" (Astbury & Leeuw, 2010, p. 364), use, and stakeholder involvement are firmly entrenched in our discipline and practice (Alkin & Taut, 2003; Brandon & Fukunaga, 2014; Daigneault et al., 2012). The origins of theory-driven evaluations can be traced back to 1930s, more specifically to Tyler's early conceptualisation of the approach (Gargani, 2003, as cited in Donaldson, 2007), but its principles to gain more prominence in the evaluation community with the publication of Chen's seminal book *Theory-driven Evaluations* in 1990 (Coryn et al., 2011). Since then, this approach gained extensive coverage in the literature, and increased popularity amongst practitioners under the guise of theory-oriented evaluations, programme theory evaluation, intervening mechanism evaluation, programme theory-driven evaluation science , and the like (Coryn et al., 2011; Donaldson, 2007). Gargani (2003, as cited in Donaldson, 2007) argues that although the practice of articulating and testing programme theory is not universally endorsed by evaluation theorists, it is widely applied by practitioners and considered as one of the preferred approaches for evaluation practice.

Similarly, concern for evaluation use can be traced back to the 1960s (Alkin & Taut, 2003), the early days of our profession (Kirkhart, 2000). As evaluators, we have a long-standing interest in the intended and unintended influence of our work, as manifested by the conceptual, symbolic or instrumental use of evaluation findings, and the learning that occurs during the evaluation process (Johnson et al., 2009). This interest is central to our professional identity (Kirkhart, 2000), so much so that the concept of use/utilisation has been the subject of extensive deliberation in theoretical writings, and is arguably the most well-researched area in the field

(Christie, 2007). Evaluators continuously strive for a greater understanding of how evaluation use can be facilitated (Alkin & Taut, 2003), and widely agree that stakeholder involvement plays a central role in this process. Stakeholder involvement is at the heart of our practice (Brandon & Fukunaga, 2013; Rodríguez-Campos, 2011), and underlies a number of formally endorsed principles for evaluators in the North, as well as firmly established and newly introduced approaches to evaluation, such as Hansen and Vedung's (2010) theory-based stakeholder evaluation approach.

Second, some of the most influential and prolific proponents of theory-driven and utilisation-focused approaches to evaluation, such as Donaldson and Patton, have a strong presence in many countries outside of the USA. For example, Donaldson's teaching has both national and global influence. His professional portfolio includes: (a) presenting workshops on evaluation theory and advanced application of programme theory in more than 25 cities in the USA, and numerous countries including SA and UK; and (b) designing and hosting a webinar series on evaluation theory, and practice challenges for evaluators working in developing countries – to date, 12,000 participants from approximately 175 countries have enrolled in this programme (Patton, 2013). Similarly, Patton's scholarly and professional work has international reach. He was the keynote speaker at the African Evaluation Association launch in 1999, and has presented at a number of international conferences, including those hosted by the UKES and the Latin American Network.

Third, it is possible that evaluators who had well-defined perspectives on evaluability were predominantly theory-driven and utilisation-focused evaluators, or at the very least strong proponents of these approaches. This claim is in no way conclusive as the theoretical orientations of participating evaluators were not explored in this study–a gap that can be filled by future research. In line with Azzam (2011), evaluators' utilisation preferences can, for example, be operationalised using selected items from Christie's (2003b) theory-to-practice instrument.

What are the implications of having shared evaluability perspectives across evaluator cohorts?

While the underlying motive for adopting a particular evaluability perspective might vary from evaluator to evaluator, the emergence of compatible perspectives across evaluator cohorts might, for example, facilitate cross-border collaboration and dialogue amongst USA and Brazil evaluators. This type of collaboration is particularly relevant in the context of cross-cultural evaluations and development evaluation (Chouinard & Cousins, 2009; Piccioto, 2003), where "a closer exchange of experiences between U.S. evaluation practitioners and their colleagues from developing countries could be mutually beneficial" (Bamberger, 2000, p.101).

In their handbook Evaluation for the 21st Century, Chelimsky and Shadish (1997, p. xii) argued that evaluation is becoming "more global and more transnational" and that "problems and programmes that we are called upon to evaluate today often extend beyond the boundaries of any one nation, any one continent, or even one hemisphere". It is therefore conceivable that an evaluability assessment team could consist of evaluators who typically practice in geographically dispersed countries. A team of this nature would be more cohesive if members shared a common understanding of what makes a programme *evaluable*. Shared task-based mental models, defined as organised knowledge structures on how a particular task is to be accomplished, have been found to significantly predict team processes, such as ease of coordination and communication, as well as team performance (Lim & Klein, 2006; Mathieu et al., 2000).

Evaluability perspectives and evaluator training.

This section addresses the third finding of this study and is framed around the discussion of why certain evaluator types were most associated with particular evaluability perspectives. While this question cannot be addressed conclusively, some tentative ideas can be proposed for validation in future research. What distinguished most evaluators with a theory-driven evaluability perspective from those with a utilisation-focused perspective is their level of formal training in evaluation. This begs the question: What is it about formal training in evaluation

(specifically university-based graduate level training) that primed highly trained evaluators in USA, UK, and SA to adopt a theory-driven perspective? Is it the curriculum orientation? Is it the set of skills imparted? Relying on curriculum outlines/ course titles alone will not allow us to draw firm conclusions about curriculum orientation or skills imparted. Future studies could adopt LaVelle and Donaldson's (2015) suggested line of inquiry to address these questions: examining the profile of those who teach evaluation courses.

The importance of pre-service training (irrespective of the curriculum orientation) in evaluation is stated in unambiguous in terms the literature (see Lavelle & Donaldson, 2015; Schwandt, 2008). Lavelle and Donaldson (2009, p. 2) argue that "evaluators are made, not born, and an extended period of training is necessary to master the evaluation-specific skills and knowledge necessary to provide quality service to clients, as well as be socialized into the professional frameworks, standards, and ethical guidelines". Are evaluators with formal training better equipped to identify evaluability challenges that are more difficult to overcome close to an evaluation? The data suggest that this might be the case.

University-based training in evaluation typically focuses on building methodological expertise (LaVelle, 2014), and developing competencies that underlie our practice, including "communicating with clients, negotiating political situations, managing team members, successfully conducting projects, capacity building, context-responsive data displays, responding to requests for proposals, and so forth" (LaVelle & Donaldson, 2015, p. 41). These training programmes are typically practitioner-focused, according to Engel, Altschuld, and Kim's (2006) findings. It is therefore reasonable to hypothesise that evaluators with a high level of formal training in evaluation are more cognisant of the inherent political and logistical challenges of our practice (i.e., the messy reality of evaluation practice) and are more open to navigate them. Bamberger, Ruth, and Mabry (2012) argue that many evaluators who practice in developed countries encounter similar challenges/constraints as their counterparts in developing countries. Evaluators with low level of formal training in evaluation, on the other hand, might perceive these challenges as beyond their control or influence. Here I am not questioning the calibre of evaluators with low level of formal training in evaluation. Many influential theorists who shaped our field are

accidental evaluators, trained in disciplines other than programme evaluation (King, 2003; Stevahn et al., 2005). There is in fact no single or well-defined path for entry into the profession, and our practice "draws from and feeds into many different contexts and domains of knowledge" (LaVelle & Donaldson, 2015, p. 41).

The data suggest that evaluators with different levels of formal training in evaluation might have different task-based mental models. LaVelle and Donaldson (2015) do not view this situation as problematic and argue that this is part of what makes our profession so interesting. Evaluators who received a high level of training in evaluation bring mastery of the evaluation process to the table, while those trained in other disciplines bring strong content and context expertise. An evaluability assessment team should ideally consist of both types of evaluators–highly trained evaluators who can communicate the value of the evaluation to resistant stakeholders and flag structural deficiencies, and content/context experts who might be able to address some of these deficiencies.

Prioritisation of Evaluability Criteria across Different Study Tasks

In this section, I discuss the results of the second research question: Are evaluators' prioritisation of evaluability criteria consistent across different study tasks?

The data suggest that evaluators' prioritisation patterns were not consistent across the different study tasks. First, the same criteria was not prioritised consistently in the same study tasks by all evaluator cohorts. Second, the relative prioritisation of evaluability criteria differed across study tasks, even within the same evaluator cohort.

Two conclusions can be inferred from the data: (a) participating evaluators were flexible in their application of evaluability criteria, that is, they did not use a formulaic approach or engage in what Tourmen (2009) calls *procedural imitation*; and (b) the contextual features of each scenario might have primed evaluators' responses.

In House's (2015) view, evaluators use both System 1 fast thinking and System 2 slow thinking when confronted with an evaluative task. This notion is based on

Kahneman's (2011) dual process model of thinking. System 1 thinking operates automatically and relies on associative memory and related learning processes (Evan, 2003). As such, its assessments of familiar and simple situations are swift, and are often accurate as they are based on knowledge accumulated through experience and practice. System 1 has "a repertoire of heuristics that enable quick judgements" (House, 2015, p. 2). In contrast, System 2 thinking is triggered when confronted with a complex and/or unfamiliar mental task that requires focused and simultaneous processing of multiple sets of information. System 2 is more measured and self-critical (Astbury, 2016) and makes use of the central working memory system (Evans, 2003). It is important to note that the tasks that activate System 1 and System 2 thinking differ across individuals (Kahneman, 2011). Multiplying 20 by 20 might be a System 1 task for some individuals but a System 2 task for others.

House (2015) argues that evaluative judgement embodies both System 1 and System 2 thinking: System 1 surveys the environment and checks its alignment with normal patterns, while System 2 deliberates on the impressions submitted by System 1 before endorsing, rejecting, or correcting them. The scenario task was an evaluative exercise where a judgment had to be made, followed by the specification of the criteria used to make this judgment. Each scenario represented a complicated evaluation situation that required careful analysis (i.e., situational recognition using System 1 and System 2 thinking) and application of practical wisdom (Hummelbrunner, 2011). The data suggest that participating evaluators, in general, might have followed this approach. Instead of applying the same criterion to inform their judgements (a form of mental trivialisation), participating evaluators seemed to have applied practical wisdom. This form of wisdom is "context dependent and operates in areas of grey, not black and white" (House, 2015, p. 80). It calls for a kind of approach that involves deliberating on what is the right thing to do "at this time, in this place, facing this situation" (Schwandt, 2003, p. 356), as opposed to implementing an algorithm or template for action. Practical wisdom is reflected in one's ability to discern the salient features/peculiarities of a situation and engage in a dialectic process that involves oscillating between the case at hand and one's repertoire of general logic (Schwandt, 2005).

The data suggest that the features of the scenarios (made salient through the manipulation of one favourable and two unfavourable evaluability dimensions) might have primed the responses of participating evaluators. The evaluability criterion that evaluators prioritised in each scenario fell under one of the evaluability dimensions that were operationalised as *unfavourable* in the scenario task. One exception was the relative prioritisation of the criterion *stakeholder willingness to collaborate with the evaluator* (along with two other criteria that relate to logistical requirements) in Scenario 2, even if the *stakeholder characteristics* dimension was *favourable* in this scenario. Do evaluators have the tendency to focus more on unfavourable conditions in their assessment of evaluability? This is a question worth addressing in future research on evaluation practice given that evaluators are faced with a wealth of potentially relevant information when they interact with stakeholders in real evaluation contexts. The nature of this information extends beyond the categories of *favourable* and *unfavourable* in real evaluation contexts. Given our processing limitations, selective attention to certain type of information is inevitable in our judgement making process (Weber & Johnson, 2009).

Evaluator Characteristics and Programme Evaluability Decisions

In this section, I discuss the results of the last research question: Do selected evaluator characteristics (practice context and experience) predict evaluators' evaluability assessments, likelihood to evaluate, and prioritisation of evaluability criteria?

The data suggest that: (a) practice context and level of experience did not predict the type of evaluability criterion prioritised in any of the scenarios; (b) evaluators practising in developed countries were more likely to characterise Scenario 1 as *evaluable with high difficulty* than as *evaluable with medium difficulty*; and (c) evaluators with low experience level (≤ 1 year to 5 years) were more likely than unlikely to evaluate the programme depicted in Scenario 1.

The first noteworthy observation is that predictive relationships were only identified in the context of Scenario 1 (i.e., the results were not replicated across all three scenarios). This particular scenario combined robust programme features with

unfavourable stakeholder characteristics and logistical requirements. It is possible that those practising in developed countries were more likely to characterise this scenario as evaluable with high difficulty (vs. medium difficulty) because they were able to make a more nuanced diagnosis of the evaluation context. Could level of experience be related to this ability? Participating evaluators who practised in developed countries were, on average, more experienced than those practising in developing countries. While level of experience did not significantly predict evaluators' assessment of evaluability in this study (the *p*-value of .052 was just over threshold for significance, suggesting that further investigations are required before conclusively accepting or rejecting the null hypothesis) there is evidence in the literature (albeit limited) that experienced and novice evaluators read and diagnose evaluation situations differently. For example, Allen (2010) found that evaluators with a high level of experience were able to accurately discern an organization's readiness for learning. In the same vein, Tourmen (2009) found that experienced evaluators interpreted the subtleties of the evaluation situation with more ease and responded to them more actively. They were able to: (a) read the implicit features/demands of the evaluation scenario (e.g., origins of the evaluation demands, compatibility of demands, attitudes towards evaluation, scale of evaluation resources, capacity and willingness of stakeholders to participate); (b) anticipate challenges that they are most likely to encounter; and (c) predict the evolution of the evaluation context. Novice evaluators, on the other hand focused more on the explicit and technical demands of the evaluation scenario, and neglected the political and situational demands of the evaluation. This could possibly explain why evaluators with low level of experience were more likely than unlikely to evaluate the programme depicted in Scenario 1. Scenario 1 was unique in that it combined robust technical features with weak political/situational features. Based on Tourmen's (2009) findings, novice evaluators tend to overlook the latter.

Study Contributions

This exploratory study is the first of its kind, in terms of area of inquiry, method and population of interest. The evaluability perspectives of four different cohorts of evaluators were derived inductively. The consistency with which these evaluators applied certain evaluability criteria across different evaluation scenarios, as well as,

the predictive relationship between selected evaluator characteristics and evaluability decisions were also explored. The methodological, theoretical and practical contributions are discussed next.

Methodological contribution.

Ramlo and Newman (2011) have recommended the use of the Q method in evaluation research and practice to derive predictor profiles, which they deem more useful and stable than individual variables. While this method has been used in programme evaluation before (e.g., Militello & Benham, 2010; Thompson, 1998; Thompson & Miller, 1983), its application is rare and sporadic. This study used an adapted version of the conventionally onerous Q Sort method, which typically uses a forced-choice condition of instruction. Instead of simply applying a previously published method to a new body of data, I adapted the method to cater for a geographically dispersed and large expert sample. Q methodological studies in programme evaluation are typically small sample investigations that use a paper-based sorting procedure. I used a self-administered electronic-based (drag-and-drop) interface that mimics the paper-based procedure. Furthermore, I used a free-sort condition of instruction and a two-stage measurement protocol to improve the reliability of the method. Whilst arguably a modest contribution, this study adds to the body of knowledge on how the Q method can be used as an investigative approach in programme evaluation.

Theoretical contribution.

In this study, I operationalised the vague and ambiguous concept of evaluability. I developed an evaluability framework for empirical validation, based on a comprehensive review of the evaluability literature. This study represents the first empirical investigation of how evaluators operationalise prescriptive theories of evaluability. It is, to the best of my knowledge, the first study on evaluation practice, which categorises the responses of four characteristically different evaluator cohorts for comparison purposes. The most notable contribution of this study is the identification of shared evaluability perspectives *across* evaluator cohorts and divergent evaluability perspectives *within* evaluator cohorts, thus adding to the

limited body of empirical knowledge on programme evaluability in four different countries.

Practical contribution.

This study generated a number of viable propositions for evaluator training, and configuration and management of evaluability assessment teams. It is worth reiterating that the conversation has to move beyond *which evaluability perspectives to transmit as part of pre-service training programmes or what is the best systematic approach for conducting an evaluability assessment*. Instead, we need to deliberate on *how to train evaluators to resolve and integrate different perspectives that might emerge in collaborative undertakings*–a plausible scenario given the results of this study. How do we transmit this type of practical wisdom, which is typically acquired through extensive experience? Trevisan (2004) identified four approaches to practical pre-service evaluation training in the literature: simulations with case descriptions, role-plays, project-focused courses, and practicum experiences. Any approach that provides pre-service exposure to the intricacies of conducting evaluability assessments in real-world settings and as part of a team is recommended here. Prototypical examples of *unevaluable* programmes or programmes *evaluable with difficulty* can be used to elicit the different evaluability perspectives that might exist within a team configured for the purpose of a simulation exercise. The same approach can be used to train prospective evaluators to read implicit political and situational challenges embedded in evaluation scenarios, and factor these in their evaluability assessments and decisions to evaluate or not. Recognising the limits of case scenarios (e.g., the difficulty to convey the full context or dynamics of the evaluation), House (2015) proposed the compilation of long stories/novels, in which events are presented as they occur (as opposed to retrospectively) and with deliberate lessons in mind. The aim is to provide a vicarious experience to evaluators or evaluators in training.

Limitations

At least four methodological limitations of this study should be highlighted. While I have provided many plausible explanations to account for the patterns that have

emerged in the data, none of them are certain. Given the exploratory nature of this study, it is difficult to explain conclusively why, for example, *theory-driven* and *utilisation-focused* evaluability perspectives emerged consistently across evaluator cohorts. This difficulty is compounded by a key study limitation, which I did not anticipate would be problematic in the context of a descriptive study of evaluation practice: Participating evaluators were not instructed to explain the reasoning behind their Q sorts or their prioritisation patterns in the scenario task.

A second limitation relates to the simulation approach used in this study. The evaluation scenarios were limited in terms of the contextual details provided. By condensing this information, it was possible to construct an instrument that could be completed within a reasonable amount of time, and thus sustain participants' attention and interest in the study. While the scenarios mimicked the nature of real evaluation situations, the absence of interaction with real life stakeholders might have prompted participants to: (a) make implicit assumptions (not known to the researcher) about the programme context, and (b) use those unknown assumptions to inform their programme evaluability decisions. Admittedly, there might be a number of extraneous factors that could have influenced participants' decision-making process, as is often the case with any simulation study. While efforts were taken to minimise such effects by using scenarios that mimicked the nature of real evaluation situations, there is simply no substitute for situations in which participants are actual actors (Christie, 2007). In other words, it is difficult to replicate conditions that unfold on the rough ground. For example, the decisions taken in real evaluation situations have tangible and enduring consequences, while those taken in simulation studies are inconsequential. Although this limitation reduces the realism of the intended simulated conditions, an argument can still be made for the use of simulation designs in systematic investigations of evaluation practice, similar to this study. Firstly, there is evidence (e.g., Cannon & Burns, 1999; O'Neil, Allred, & Dennis, 1997) that results derived from simulation studies are comparable to those of other methodological approaches. Secondly, the scenario task was undeniably a useful tool for developing a preliminary understanding of the effects of different evaluability criteria on programme evaluability decisions. The results of this study should however be considered in light of the inherent limitations of simulation

designs, and any attempts to generalise the results from this study must be undertaken with caution.

It is worth reiterating that this study, like any study that relies predominantly on purposive and snowball sampling strategies, carries the risk of selection bias. As such, the extent to which the findings are generalisable to other evaluation contexts or to evaluators who did not self-select into this study remains uncertain. There is a possibility that participating evaluators were inherently different from those who declined participation or withdrew from the study. The direction and magnitude of non-response bias could not be established in this study. It is also entirely possible that evaluators with certain evaluability paradigms were underrepresented in this study, given the type of sampling strategy used and the low response rate from the UK and SA cohorts. In retrospect, the exclusion criteria should have been more conservative in order to address the conceptual distinction between *country of residence* and *country of practice*, with the latter being more salient in the context of this study. While the inclusion of the item: *where do you mostly do evaluation work?* served to distinguish evaluators who practiced in developing countries from those who practiced in developed countries, instances where an evaluator resided in a developing country but mostly practiced in a developed country or vice-versa, were not accounted for in the first set of analyses. However, such instances were negligible (see Table E1 in Appendices), and as such unlikely to skew the results. A notable exception was for the UK cohort, with 40% of participating evaluators practicing in developing countries. Nonetheless, it is worth reiterating that no firm conclusions could be/were drawn for this particular cohort due to the low response rate.

In addition, while the statistical analyses used in this study are appropriate for an exploratory investigation, they have certain inherent limitations: some in terms of the nature of the conclusions that can be drawn, and others in terms of the process that was used to arrive at those conclusions. For instance, the patterns identified in the CA are not generalisable beyond the study sample, and one should keep in mind that the two-dimensional CA maps used to isolate those patterns are not perfect

representations of the association between different evaluability criteria and the various study tasks.

For the MLR, I collapsed participants' assessments of evaluability and their likelihood of conducting an evaluation, initially measured on a 10-point scale, into three categories (Low, Medium, and High). This approach was used to simplify the analysis and the interpretation of results. One could, however, question the assumptions underlying the three categories, and their associated ranges (e.g., 1-4 = High; see Table 23). There are no recognised cut-off points for the outcome variables of interest in the literature. I selected the most intuitive categorisation. This was deemed preferable to performing several analyses and choosing optimal ranges that produce significant but possibly spurious results. In retrospect, using this approach for the sake of simplicity might have compromised the precision of the data. Categorising a continuous variable, regardless of the method used (e.g., a binary/median split), implies loss of information and statistical power (Altman & Royston, 2006; Royston, Altman, & Sauerbrei, 2006). In other words, it becomes harder to detect effects that are in fact present (in this case the predictive relationship between selected evaluator characteristics and programme evaluability decisions). Categorisation of a continuous variable might also increase the risk of Type 1 errors (Austin & Brunner, 2004). I tried to offset this inflated risk by using a more stringent significance level ($p < .025$) but the wide confidence intervals indicate the need to replicate the study with more data points and more precise measurement for this particular analysis.

Conclusions and Directions for Future Research

Although predominantly descriptive, this study provides valuable insight into the evaluation practice of four different cohorts of evaluators, more specifically how they operationalise prescriptive theories of programme evaluability. As Smith (1993, p. 240) remarked over two decades ago, "if evaluation theories cannot be uniquely operationalised, then empirical tests of their utility becomes increasingly difficult". The results of this study suggest that there is no algorithm for evaluability, and that perhaps decisions about evaluability are the outcome of a discourse between evaluators with different perspectives, and programme stakeholders. We now have

preliminary evidence that: (a) evaluators do not share a unified perspective towards evaluability; (b) perspectives on evaluability are shaped, in part, by level of formal training; (c) evaluators do not adopt a formulaic approach to assess programme evaluability; and (d) stakeholders' willingness to collaborate with the evaluator is consistently prioritised over other stakeholder characteristics such as transparency and authority. The next step would be to unpack the *why*. A suggested approach would be to replicate this study on a smaller scale, using Tourmen's (2009) work analysis method, and descriptively map out the back and forth movements, the difficulties and compromises involved in programme evaluability decision-making. We somehow need to restore the complexity of real evaluation situations in subsequent studies of evaluation practice and capture assessment of programme evaluability in action. More specifically we need to capture evaluators' application of System 1 and System 2 thinking as it unfolds on the rough ground. This will allow us to have a more refined understanding of the logic in use/working logic in programme evaluability.

We also need to restore the reality of evaluability assessments in subsequent investigations. Many evaluability assessments are collaborative undertakings. In this study participating evaluators reflected on their individual practice and reported on the evaluability criteria that they prioritised. Future research could examine the extent of alignment between individual and team level prioritisation patterns, and associated implications on collaborative work and the evaluability assessment process. It is conceivable that evaluators with divergent perspectives on evaluability might find it difficult to work collaboratively.

Another avenue for future research would be to further unpack a secondary finding of this study: stakeholders' willingness to collaborate with the evaluator, adequacy of evaluation timeframe, and clear specification of programme goals were most frequently prioritised (in absolute terms) across all evaluator cohorts. For instance, we need to understand the extent of stakeholder collaboration that evaluators perceive as optimal during an evaluability assessment. It makes intuitive sense to seek stakeholder collaboration in any endeavour but the expected/desired depth and breadth of that collaboration might vary depending on the evaluation context or the characteristics of the evaluator or evaluation team (Fitzpatrick, 2004). A number of

pertinent questions come to mind: In a complex evaluation situation with multiple evaluation sites and stakeholder groups, what level of collaboration is desired from each stakeholder group? How do evaluators operationalise stakeholder collaboration in an evaluability assessment? Does this operationalisation vary across evaluators and practice contexts? This study has generated many more questions than answers. These questions, in my view, define an agenda for research in the area of programme evaluability– a previously unexplored, yet fundamental area in our evaluation practice.

References

Abrahams, M. A. (2015). A review of the growth of monitoring and evaluation in South Africa: Monitoring and evaluation as a profession, an industry and a governance tool. *African Evaluation Journal, 3*(1), 1-8.

African Development Bank. (2008). *South Africa economic outlook*. Retrieved from http://www.afdb.org/en/

African Evaluation Association. (2002). *African evaluation guidelines*. Retrieved from http://www.afrea.org/

Alberti, G. (2013). An R script to facilitate correspondence analysis: A guide to the use and the interpretation of results from an archaeological perspective. *Archeologia e Calcolatori*, *24*, 25–53. Retrieved from http://soi.cnr.it/archcalc/

Alkin, M. C. (Ed.). (2004). Evaluation roots: Tracing theorists' views and influences. Thousand Oaks, California: Sage Publications, Inc.

Alkin, M. C., & Taut, S. M. (2003). Unbundling evaluation use. *Studies in Educational Evaluation*, *29*(1), 1-12.

Alkin, M. C., & Christie, C. A. (2004). An evaluation theory tree. In M. C Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 12-65). Thousand Oaks, California: Sage Publications, Inc.

Alkin, M. C., & Christie, C. A. (2005). Unraveling theorists' evaluation reality. *New Directions for Evaluation*, *2005*(106), 111–128.

Allen, M. (2010). *Steeping the organization's tea: Examining the relationship between evaluation use, organizational context, and evaluator characteristics* (Doctoral dissertation, Case Western Reserve University, Ohio). Retrieved from http://msass.case.edu/downloads/vgroza/Allen_Dissertation.pdf

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj*, *332*, 1080. Abstract retrieved from http://www.bmj.com/content/332/7549/1080.1

Altschuld, J. W., Engle, M., Cullen, C., Kim, I., & Rae Macce, B. (1994). The 1994 directory of evaluation training programs. *New Directions for Program Evaluation, 1994*(62), 71–94.

Altschuld, J. W., Yoon, J. S., & Cullen, C. (1993). The utilization of needs assessment results. *Evaluation and Program Planning*, *16*(4), 279–285.

American Evaluation Association. (2012). *Benefits of AEA membership*. Retrieved from http://www.eval.org/

Arretche, M. (2002). Federalism and inter-governmental relations in Brazil: social-program reforms. *Dados*, *45*(3), 431–458. Retrieved from http://www.scielo.br/scielo.php?script=sci_serial&pid=0011-5258&lng=en&nrm=iso

Astbury, B. (2016). Reframing how evaluators think and act: New insights from Ernest House. *Evaluation*, *22*(1), 58–71.

Astbury, B., & Leeuw, F. L. (2010). Unpacking black boxes: Mechanisms and theory building in evaluation. *American Journal of Evaluation*, *31*(3), 363–381.

Austin, P. C., & Brunner, L. J. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in medicine, 23*(7), 1159-1178.

Ayers, T. D. (1987). Stakeholders as partners in evaluation: A stakeholder-collaborative approach. *Evaluation and Program Planning*, *10*(3), 263–271.

Azzam, T. (2011). Evaluator characteristics and methodological choice. *American Journal of Evaluation*, *32*(3), 376–391.

Azzam, T., & Szanyi, M. (2011). Designing evaluations: A study examining preferred evaluation designs of educational evaluators. *Studies in Educational Evaluation*, *37*(2), 134–143.

Babbie, E., & Mouton, J. (2001). *The practice of social science research.* Belmont, CA: Wadsworth.

Bamberger, M. (2000). The evaluation of international development programs: A view from the front. *American Journal of Evaluation*, *21*(1), 95–102.

Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation*, *25*(1), 5–37.

Bamberger, M., Rugh, J., & Mabry, L. (2012). *Real world evaluation: Working under budget, time, data, and political constraints*. Thousand Oaks, California: Sage Publications, Inc.

Bayaga, A. (2010). Multinomial logistic regression: Usage and application in risk analysis. *Journal of Applied Quantitative Methods*, *5*(2), 288–297. Retrieved from http://www.jaqm.ro/

Beh, E. J., Lombardo, R., & Simonetti, B. (2011). A European perception of food using two methods of correspondence analysis. *Food Quality and Preference*, *22*(2), 226–231.

Bendixen, M. (1996). A practical guide to the use of correspondence analysis in marketing research. *Marketing Research On-Line, 1*(1), 16-36.

Bertelsmann, S. (2016). *Brazil Country Report*. Retrieved from http://www.bti-project.org/en/reports/country-reports/detail/itc/BRA/

Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation*, *1987*(33), 5–18.

Bogetić, Z. and Fedderk, J. W. (2006). *International benchmarking of South Africa's infrastructure performance* (World Bank Policy Research Working Paper 3830). Retrieved from Republic of South African Presidency website: http://www.thepresidency.gov.za/electronicreport/downloads/volume_4/business_case_viability/BC1_Research_Material/International_Benchmarking_of_SAs_Infrastructure_Performance.pdf

Borel, F. (2015). Closing South Africa's high-skilled worker gap: Higher education challenges and pathways (African Economic Brief 6-7). Retrieved from The African Development Bank website: http://www.afdb.org/fileadmin/uploads/afdb/Documents/Knowledge/AEB_Vol_6_i_7_-Closing_South_Africas_High-Skilled_Worker_Gap__Higher_Education_Challenges_and_Pathways.pdf

Boyer, J. F., & Langbein, L. I. (1991). Factors influencing the use of health evaluation research in Congress. *Evaluation Review*, *15*(5), 507–532.

Brandon, P. R., & Fukunaga, L. L. (2014). The state of the empirical research literature on stakeholder involvement in program evaluation. *American Journal of Evaluation*, *35*(1), 26-44.

Braskamp, L. A., Brown, R. D., & Newman, D. L. (1982). Studying evaluation utilization through simulations. *Evaluation Review*, *6*(1), 114–126.

Brisolara, S. (1998). The history of participatory evaluation and current debates in the field. *New Directions for Evaluation*, *1998*(80), 25–41.

Brown, S. R. (1980). *Political subjectivity: Applications of Q methodology in political science*. New Haven, CT: Yale University Press.

Brown, S. R. (1993). A primer on Q methodology. *Operant subjectivity, 16*(3/4), 91-138.

Brunner, R. D. (2006). A paradigm for practice. *Policy Sciences*, *39*(2), 135-167.

Bruns, B., Evans, D., & Luque, J. (2012). *Achieving world-class education in Brazil: The next agenda*. Washington DC: World Bank Publications.

Burdescu, R., del Villar, A., Mackay, K., Rojas, F., & Saavedra, J. (2005). *Institutionalizing monitoring and evaluation systems: Five experiences from Latin America* (World Bank En Breve No. 78-34642). Retrieved from The World Bank website: https://openknowledge.worldbank.org/bitstream/handle/10986/10322/346420ENGLISH0en0breve078SEP05M1EEN.pdf?sequence=1

Burns, R. P., & Burns, R. (2008). *Business research methods and statistics using SPSS*. London: Sage Publications Ltd.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24*(4), 409–429.

Cannon, H. M., & Burns, A. C. (1999). A framework for assessing the competencies reflected in simulation performance. *Developments in Business Simulation and Experiential Learning*, *26*, 40-44.

Caracelli, V. J. (2000). Evaluation use at the threshold of the twenty-first century. *New directions for evaluation, 2000*(88), 99-111.

Chelimsky, E. (2006). The purposes of evaluation in a democratic society. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.), *The SAGE handbook of evaluation* (pp. 33–55). Thousand Oaks, CA: SAGE Publications.

Chelimsky, E. (2013). Balancing evaluation theory and practice in the real world. *American Journal of Evaluation*, *34*(1), 91–98.

Chelimsky, E., & Shadish, W. R. (Eds.). (1997). *Evaluation for the 21st century: A handbook*. Thousand Oaks, CA: Sage Publications.

Chen, H. T. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage Publications, Inc.

Chen, H. T. (2005). A conceptual framework of programme theory for practitioners. In L. Lech (Ed.), *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness* (pp. 15–43). Thousand Oaks, CA: SAGE Publications.

Chen, H. T., & Rossi, P. H. (1989). Issues in the theory-driven perspective. *Evaluation and Program Planning*, *12*(4), 299–306.

Chianca, T. (2007). An update on evaluation in the Latin American and Caribbean region. *Journal of Multidisciplinary Evaluation*, *2*(2), 137-144.

Chouinard, J. A., & Cousins, J. B. (2009). A review and synthesis of current research on cross-cultural evaluation. *American Journal of Evaluation*, *30*(4), 457–494.

Christie, C. A. (2003a). Understanding evaluation theory and its role in guiding practice: Formal, folk, and otherwise. *New Directions for Evaluation*, *2003*(97), 91–93.

Christie, C. A. (2003b). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. *New Directions for Evaluation*, *2003*(97), 7–36.

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions an empirical examination. *American Journal of Evaluation*, *28*(1), 8–25.

Christie, C. A. (2011). Advancing empirical scholarship to further develop evaluation theory and practice. *The Canadian Journal of Program Evaluation*, *26*(1), 1–18. Retrieved from http://evaluationcanada.ca/canadian-journal-program-evaluation

Christie, C. A., & Alkin, M. C. (2008). Evaluation theory tree re-examined. *Studies in Educational Evaluation, 34*(3), 131-135.

Christie, C. A., & Azzam, T. (2005). What theorists say they do: A brief description of theorists' approaches. *New Directions for Evaluation*, *2005*(106), 15–26.

Cloete, F. (2009). Evidence-based policy analysis in South Africa: Critical assessment of the emerging government-wide monitoring and evaluation system. *Journal of Public Administration, 44*(2), 293–311. Retrieved from http://paperroom.ipsa.org/

Cloete, F., Rabie, B. & De Coining, C. (Eds.). (2014). *Evaluation management in South Africa and Africa*. Stellenbosch: SUN PRESS Imprint.

Codato, A. N. (Ed.). (2006). *Political transition and democratic consolidation: Studies on contemporary Brazil*. Hauppauge, NY: Nova Publishers.

Cooper, R. (2014). *Psychiatry and philosophy of science*. New York, NY: Routledge.

Coryn, C. L., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation*, *32*(2), 199–226.

Coryn, C. L., Ozeki, S., Wilson, L. N., Greenman, G. D., Schröter, D. C., Hobson, K. Azzam, T., Vo, A. T. (2016). Does research on evaluation matter? Findings from a survey of American Evaluation Association members and prominent evaluation theorists and scholars. American Journal of Evaluation, 37(2), 159-173.

Cousins, J. B., & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, *14*(4), 397–418.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of educational research*, *56*(3), 331-364.

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. *New Directions for Evaluation*, *1998*(80), 5–23.

Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social science research.* UK: Lawrence Erlbaum and Associates.

Cross, R. M. (2005). Exploring attitudes: the case for Q methodology. *Health education research, 20*(2), 206-213.

Dahler-Larsen, P., & Schwandt, T. A. (2012). Political culture as context for evaluation. *New Directions for Evaluation, 2012*(135), 75–87.

Daigneault, P. M., Jacob, S., & Tremblay, J. (2012). Measuring stakeholder participation in evaluation: An empirical validation of the Participatory Evaluation Measurement Instrument (PEMI). *Evaluation Review*, *36*(4), 243-271.

Davies, R. (2008). *M&E training providers*. Retrieved from MandE NEWS website: http://www.mande.co.uk/training.htm#United%20Kingdom

Davies, R. (2013). *Planning evaluability assessment: A synthesis of the literature with recommendations* (Working Paper no. 40). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/248656/wp40-planning-eval-assessments.pdf

Davies, R., & Payne, L. (2015). Evaluability assessments: Reflections on a review of the literature. *Evaluation*, *21*(2), 216–231.

Demarteau, M. (2002). A theoretical framework and grid for analysis of programme-evaluation practices. *Evaluation*, *8*(4), 454–473.

Department for International Development. (2015). *Statistics on international development* (Report 1). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/482322/SID2015c.pdfhttps://www.bond.org.uk/data/files/ICAIs-Approach-to-Effectiveness-and-VFM.pdf

Department of Planning, Monitoring and Evaluation (2016). *National Evaluation Plan 2016-17 to 2018-19*. Retrieved from http://www.dpme.gov.za/keyfocusareas/evaluationsSite/Evaluations/National%20Evaluation%20Plan%202016-17%2016.03.31.pdf

Derlien, H. (1990). Genesis and structure of evaluation efforts in comparative perspective. In R. C. Rist (Ed.), *Program evaluation and the management of government: Patterns and prospects across eight nations* (pp. 146–176). New Brunswick, NJ: Transaction Publishers.

Doey, L., & Kurta, J. (2011). Correspondence analysis applied to psychological research. *Tutorials in Quantitative Methods for Psychology, 7*(1), 5-14.

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications*. New York, NY: Routledge.

Donaldson, S. I., Gooler, L. E., & Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. *American Journal of Evaluation, 23*(3), 261–273.

Donaldson, S. I., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: Developing practical knowledge. The handbook of evaluation: Policies, programs, and practices, 56-75.

Du Plessis, T. C. (2005). A theoretical framework of corporate online communication" A marketing public relations perspective (Doctoral dissertation, University of South Africa). Retrieved from http://uir.unisa.ac.za/handle/10500/2271

Duffy, F. M. (2009). Paradigms, mental models, and mindsets: Triple barriers to transformational change in school systems. *Educational Technology, 54*(3), 29–33. Retrieved from http://www.bookstoread.com/etp

Dziopa, F., & Ahern, K. (2011). A systematic literature review of the applications of Q-technique and its methodology. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 7*(2), 39.

Eghbalighazijahani, M. A., Hine, J., & Kashyap, A. (2013, September). How to do a better Q-methodological research: A neural network method for more targeted decision making about the factors influencing Q-study. *Proceedings of ITRN2013*, Dublin. Retrieved from http://www.itrn.ie/uploads/1287_Eghbalighazijahani[1].pdf

Engela, R., & Ajam, T. (2010). *Implementing a government-wide monitoring and evaluation system in South Africa.* Washington DC: The World Bank.

Engle, M., Altschuld, J. W., & Kim, Y. C. (2006). 2002 survey of evaluation preparation programs in universities an update of the 1992 American evaluation association–sponsored study. *American Journal of Evaluation*, *27*(3), 353–359.

Epstein, D., & Klerman, J. A. (2012). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. *Evaluation Review*, *36*(5), 375–401.

Evans, D., & Kosec, K. (2012). *Early child education: Making programs work for Brazil's most important generation*. Washington DC: The World Bank.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454-459.

Feinstein, O. N. (2002). Use of evaluations and the evaluation of their use. *Evaluation*, *8*(4), 433–439.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage Publications Ltd.

Finckenauer, J. O., Margaryan, S., & Sullivan, M. L. (2005). Evaluability assessment in juvenile justice a case example. *Youth Violence and Juvenile Justice*, *3*(3), 265–275.

Firme, T. P., Letichevsky, A. C., Dannemann, Â. C., & Stone, V. (2009). Evaluation culture and evaluation policy as guides to practice: Reflections on the Brazilian experience. *Ensaio: Avaliação e Políticas Públicas em Educação*, *17*(62), 169–180. Retrieved from http://www.scielo.br/

Fitzpatrick, J. L. (2004). Exemplars as case studies: Reflections on the links between theory, practice, and context. *American Journal of Evaluation, 25*(4), 541-559.

Fitzpatrick, J. L. (2012). An introduction to context and its role in evaluation practice. In D. J. Rog, J. L. Fizpatrick & R. F. Conner (Eds.), *Context: A framework for its influence on evaluation practice* (pp. 7–24). San Francisco, CA: Wiley Subscription Services Inc.

Fitzpatrick, J., Christie, C., & Mark, M. M. (Eds.). (2009). *Evaluation in action: Interviews with expert evaluators*. Thousand Oaks, CA: SAGE Publications.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation*, *1995*(68), 15–32.

Glynn, D. (2014). Correspondence Analysis. An exploratory technique for identifying usage patterns. In D. Glynn & J. Robinson (Eds.). *Corpus Methods for Semantics. Quantitative Studies in Polysemy and Synonymy* (pp.443-486). Retrieved from https://www.researchgate.net/publication/275647208_Correspondence_analysis_Exploring_data_and_identifying_patterns

Grasso, P. G. (1996). End of an era: Closing the U.S. general accounting office's program evaluation and methodology division. *Evaluation Practice, 17*(2), 115–132. Retrieved from http://www.sciencedirect.com/science/journal/08861633

Grasso, P. G. (2003). What makes an evaluation useful? Reflections from experience in large organizations. American Journal of Evaluation, 24(4), 507-514.

Gray, A., & Jenkins, B. (1982). *Policy analysis in British central government: the experience of PAR. Public Administration*, 60(4), 429-450.

Gray, A., & Jenkins, B. (2002). Policy and program evaluation in the United-Kingdom: A reflective state. In J. E. Furubo, R. C. Rist & R. Sandahl (Eds.), *International atlas of evaluation* (pp. 129–153). London: Transaction Publishers.

Greenacre, M. J. (1989). *Theory and applications of correspondence analysis*. London, England: Academic Press Limited.

Greenacre, M. (2007). *Correspondence analysis in practice*. Florida: Chapman & Hall/CRC press.

Greenacre, M. J. (2010). *Biplots in practice*. Madrid, Spain: Fundacion BBVA.

Greenacre, M. (October, 2011). *The contributions of rare objects in correspondence analysis*. Paper presented at 1st International Workshop of CARME, Assos, Turkey. Retrieved from http://repositori.upf.edu/bitstream/handle/10230/19864/1278.pdf?sequence=1

Greene, J. C. (2005a). *Context*. In S. Mathison (Ed.), Encyclopedia of evaluation (pp. 82–84). Thousand Oaks, CA: Sage

Greene, J. C. (2005b). Evaluators as stewards of the public good. In S. Hood, R. Hopson & H. Frierson (Eds.), *The role of culture and cultural context: A mandate for inclusion, the discovery of truth, and understanding in evaluative theory and practice* (pp. 7–20). Scottsdale, AZ: Information Age Publishing Inc.

Greene, J. C. (2006). Method choice: Five discussant commentaries. *New Directions for Evaluation, 2007(*113), 111–127.

Guimaraes, T. B., & Campos, E. (2010). Monitoring and evaluation system in the Minas Gerais state government: Aspects of management. In G. Acevedo, K. Rivera, L. Lima, & H. Hwang (Eds.), *Challenges in monitoring and evaluation: an opportunity to institutionalize M&E systems*. Washington, DC: The International Bank for Reconstruction and Development/The World Bank.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. Upper Saddle River, NJ: Pearson International.

Hansen, H. F. (2005). Choosing evaluation models: A discussion on evaluation design. *Evaluation*, *11*(4), 447–462.

Hansen, M. B., & Vedung, E. (2010). Theory-based stakeholder evaluation. *American Journal of Evaluation*, *31*(3), 295–313.

Henriques, A., Pinho, J., Azevedo, J. P., & Newman, J. L. (2010). The Brazilian Monitoring and Evaluation Network: A report on the creation and development process. In G. Acevedo, K. Rivera, L. Lima, & H. Hwang (Eds.), *Challenges in monitoring and evaluation: an opportunity to institutionalize M&E systems*. Washington, DC: The International Bank for Reconstruction and Development/The World Bank.

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. *New Directions for Evaluation*, *2003*(97), 69–80.

HM Treasury. (2011). The Green Book: Appraisal and evaluation in central government. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/220541/green_book_complete.pdf

Hoffman, D. L., & De Leeuw, J. (1992). Interpreting multiple correspondence analysis as a multidimensional scaling method. *Marketing Letters, 3*(3), 259-272.

Hood, C., James, O., Jones, G., Scott, C., & Travers, T. (1998). Regulation inside government: where new public management meets the audit explosion. *Public Money and Management, 18*(2), 61-68.

Hood, C., James, O., & Scott, C. (2000). Regulation of government: has it increased, is it increasing, should it be diminished?. *Public Administration, 78*(2), 283-304.

Horst, P., Nay, J. N., Scanlon, J. W., & Wholey, J. S. (1974). Program management and the federal evaluator. *Public Administration Review, 34*(4), 300–308. Retrieved from http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1540-6210

House, E. R. (2015). *Evaluating: Values, biases and practical wisdom*. Charlotte, NC: Information Age Publishing.

Hummelbrunner, R. (2011). Systems thinking and evaluation. *Evaluation*, *17*(4), 395–403.

Independent Commission for Aid Impact. (2011). *ICAI's approach to effectiveness and value for money* (Report 1). Retrieved from https://www.bond.org.uk/data/files/ICAIs-Approach-to-Effectiveness-and-VFM.pdf

International Labour Organization (2012). *Using the evaluability assessment tool* (Guidance Note 11). Retrieved from ILO website: http://www.ilo.org/wcmsp5/groups/public/---ed_mas/---eval/documents/publication/wcms_165984.pdf

International Organization for Cooperation in Evaluation (2015). *BMEN member profile details*. Retrieved from http://ioce-vopes.wildapricot.org/widget/Sys/PublicProfile/26039999/3594433

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use a review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, *30*(3), 377–410.

Joyce, P. G. (2011). The Obama Administration and PBB: Building on the Legacy of Federal Performance - Informed Budgeting? *Public Administration Review, 71*(3), 356-367.

Jung, S. M., & Schubert, J. G. (1983). Evaluability assessment: A two-year retrospective. *Educational Evaluation and Policy Analysis*, *5*(4), 435–444.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kalley, J. A., Schoeman, E., & Andor, L. E. (1999). *Southern African political history: a chronology of key political events from independence to mid-1997*. Westport: Greenwood Publishing Group.

Kaufman-Levy, D., & Poulin, M. (2003). *Evaluability assessment: Examining the readiness of a program for evaluation*. Washington DC: Juvenile Justice Evaluation Center, Justice Research and Statistics Association. Retrieved from http://www.jrsa.org/pubs/juvjustice/evaluability-assessment.pdf

Kelly, J. (2003). The audit commission: Guiding, steering and regulating local government. *Public Administration, 81*(3), 459–476.

Kennedy, R., Riquier, C., & Sharp, B. (1996). Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting Measurement and Analysis for Marketing, 5*, 56-70.

Kierzenkowski, R., Pain, N., Rusticelli, E., and Zwart, S. (2016). *The economic consequences of Brexit: A taxing decision*. Retrieved from http://www.oecd.org/unitedkingdom/The-Economic-consequences-of-Brexit-27-april-2016.pdf

King, J. A. (1998). Making sense of participatory evaluation practice. *New directions for evaluation*, (80), 57-67.

King, J. A. (2003). The challenge of studying evaluation theory. *New Directions for Evaluation, 2003*(97), 57–68.

King, J. A., Stevahn, L., Ghere, G., & Minnema, J. (2001). Toward a taxonomy of essential evaluator competencies. *American Journal of Evaluation*, *22*(2), 229–247.

King, J., & Greenseid, L. (2007). The Oral History of Evaluation, Part 5: An Interview with Michael Quinn Patton. *American Journal of Evaluation, 28*(1), 102-114.

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation*, *2000*(88), 5–23.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management*, *20*(2), 403–437.

Kundin, D. M. (2008). *Everyday approaches to evaluation: A study of how evaluators make practice decisions* (Doctoral dissertation, University of Minnesota). Retrieved from http://search.proquest.com.ezproxy.uct.ac.za/docview/304532263

Kundin, D. M. (2010). A conceptual framework for how evaluators make everyday practice decisions. *American Journal of Evaluation*, *31*(3), 347–362.

LaVelle, J. M. (2014). *An examination of evaluation education programs and evaluator skills across the world* (Doctoral dissertation, Claremont Graduate University). Retrieved from http://search.proquest.com/docview/1527108167

LaVelle, J. M., & Donaldson, S. I. (2010). University-based evaluation training programs in the United States 1980–2008: An empirical examination. *American Journal of Evaluation, 31*(1), 9–23.

LaVelle, J. M., & Donaldson, S. I. (2015). The state of preparing evaluators. *New Directions for Evaluation*, *2015*(145), 39–52. doi:10.1002/ev.20110

Leeuw, F. L. (2011). On the contemporary history of experimental evaluations and its relevance for policy making. In O. Rieper, F. L. Leeuw, T. Ling (Eds.), *Concepts, generation and use of evidence: Comparative policy evaluation* (pp. 11-26). New Jersey: Transaction Publishers.

Levesque, L. L., Wilson, J. M., & Wholey, D. R. (2001). Cognitive divergence and shared mental models in software development project teams. *Journal of Organizational Behavior*, *22*(2), 135–144.

Leviton, L. C., Khan, L. K., Rog, D., Dawkins, N., & Cotton, D. (2010). Evaluability assessment to improve public health policies, programs, and practices. *Annual Review of Public Health, 31*, 213–233. doi:10.1146/annurev.publhealth.012809.103625

Lim, B. C., & Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior*, *27*(4), 403–418.

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, *14*(5), 331–352.

Lopez-Calva, L. F., & Rocha, S. (2012). *Exiting Belindia? Lesson from the recent decline in income inequality in Brazil*. Retrieved from https://wdronline.worldbank.com/bitstream/handle/10986/12808/701550ESW0P1230IC00InequalityBrazil.pdf?sequence=1

Louw, J. (1995). Evaluation in South Africa: An introduction. *Evaluation and Program Planning, 18*(4), 351–353.

Madzivhandila, T. P. (2010). *A practical programme evaluation model for the Limpopo department of agriculture*. (Doctoral thesis, University of New England, Australia). Retrieved from https://e-publications.une.edu.au/vital/access/manager/Repository/une:11479

Maia, J., Mondi, L., & Roberts, S. (2005). *Industrial development and industrial finance in Brazil and South Africa: Comparative assessment*. Paper presented at the Annual Forum 2005: Trade and Uneven Development: Opportunities and Challenges, Gauteng.

Martin, R., Gardiner, B., Tyler, P (2014). The evolving economic performance of UK cities: city growth patterns 1981-2011. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/358326/14-803-evolving-economic-performance-of-cities.pdf

Martin, S. (2002). The modernization of UK local government: Markets, managers, monitors and mixed fortunes. *Public Management Review*, *4*(3), 291–307.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*(2), 273–283.

McGuire, M., & Zorzi, R. (2005). Evaluator competencies and performance development. *Canadian Journal of Program Evaluation, 20*(2), 73.

Melkers, J., & Roessner, D. (1997). Politics and the political setting as an influence on evaluation activities: National research and technology policy programs in the United States and Canada. *Evaluation and Program Planning, 20*(1), 57–75.

Mertens, D. M., & Wilson, A. T. (2012). *Program evaluation theory and practice: A comprehensive guide*. New York, NY: Guilford Press.

Militello, M., & Benham, M. K. (2010). “Sorting Out” collective leadership: How Q-methodology can be used to evaluate leadership development. *The Leadership Quarterly*, *21*(4), 620-632.

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation*, *31*(3), 390–399. Retrieved from http://aje.sagepub.com/

Mohanty, R., Thompson, L., & Coelho, V. S. (2011). Mobilising the state? Social mobilisation and state interaction in India, Brazil and South Africa. *IDS Working Papers*, *2011*(359), 1–39. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.2040-0209.2011.00359_2.x/pdf

Molden, D. C. (2014). Understanding priming effects in social psychology: What is social priming and how does it occur? Social Cognition, 32, 243–249.

Mouton, C. (2010). *The history of programme evaluation in South Africa* (Doctoral thesis, University of Stellenbosch, Stellenbosch, South Africa). Retrieved from https://www.google.co.za/search?q=The+history+of+programme+evaluation+in+South+Africa&oq=The+history+of+programme+evaluation+in+South+Africa&aqs=chrome..69i57j69i64l2.305j0j7&sourceid=chrome&ie=UTF-8

Mullard, M. (2001). New Labour, new public expenditure: The case of cake tomorrow. *The Political Quarterly*, *72*(3), 310–321.

Nay, J. N., & Kay, P. (1982). *Government oversight and evaluability assessment: Its always more expensive when the carpenter types*. New York, NY: Lexington Books.

Nayyar, D. (2008). *China, India, Brazil and South Africa in the world economy: Engines of growth?* Paper presented at the meeting of UNU-WIDER World Institute for Development Economics, Jawaharlal Nehru University, New Delhi.

Neirotti, N. (2012). Evaluation in Latin America: Paradigms and practices. *New Directions for Evaluation*, *2012*(134), 7–16.

Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two-and three-dimensional graphics: The ca package. *Journal of Statistical Software, 20*(3), 1-13.

Neri, M. C., & Buchmann, G. (2007). *The Brazilian education quality index: Measurement and incentives upgrades* (Technical Report *No.* 686). Retrieved from: http://www.cps.fgv.br/ibrecps/discussao/EE2008_QualiEduc_Paper_International_Submission_Final.pdf

Nevo, D. (1982). The international context for research on evaluation. *American Journal of Evaluation*, *3*(4), 73–75.

Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (Eds.). (2010). *Handbook of practical program evaluation* (3rd ed.). New York, NY: John Wiley & Sons.

Nielsen, L. (2011). *Classifications of countries based on their level of development: How it is done and how it could be done* (IMF Working Papers No. 11/31). Washington, DC: IMF.

OECD (2014). *Social expenditure update* .Retrieved from https://www.oecd.org/els/soc/OECD2014-Social-Expenditure-Update-Nov2014-8pages.pdf

OECD (2015a).*Education at a glance 2015: How does the United States compare?* Retrieved from http://www.keepeek.com/Digital-Asset-Management/oecd/education/education-at-a-glance-2015/united-states_eag-2015-86-en#.V8FThZh9600

OECD (2015b).*Health at a Glance 2015: How does the United Kingdom compare?* Retrieved from http://www.oecd.org/unitedkingdom/Health-at-a-Glance-2015-Key-Findings-UK.pdf

OECD (2015b).*Health at a Glance 2015: How does the United States compare?* Retrieved from http://www.oecd.org/unitedstates/Health-at-a-Glance-2015-Key-Findings-UNITED-STATES.pdf

OECD (2015c).*Education at a Glance 2015: OECD indicators*. Retrieved from http://www.keepeek.com/Digital-Asset-Management/oecd/education/education-at-a-glance-2015/united-kingdom_eag-2015-85-en#.V8GTDph97IU#page1

OECD (2016a).*How's life in the United States*. Retrieved from http://www.oecd.org/unitedstates/Better-Life-Initiative-country-note-United-States.pdf

OECD (2016b). *Employment Outlook 2016: How does the United States compare?* Retrieved from http://www.oecd.org/unitedstates/Employment-Outlook-UnitedStates.pdf

OECD (2016c).*Social expenditure database: Aggregated data*. Retrieved from https://stats.oecd.org/Index.aspx?DataSetCode=SOCX_AGG

OECD (2016d). *How's life in the United Kingdom.* Retrieved from http://www.oecd.org/unitedkingdom/Better-Life-Initiative-country-note-United-Kingdom.pdf

OECD (2016e). *Employment outlook 2016: How does the United Kingdom compare?* Retrieved from http://www.oecd.org/unitedkingdom/Employment-Outlook-UnitedKingdom-EN.pdf

Office of Management and Budget. (2009). Increased emphasis on programme evaluations (M-10-01). Retrieved from https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-01.pdf

O'Neil Jr, H. F., Allred, K., & Dennis, R. A. (1997). Use of computer simulation for assessing the interpersonal skill of negotiation. In H.F O'Neil Jr & National Centre for Research on Evaluation, Standards, and Student Testing (Eds.), *Workforce readiness: Competencies and assessment* (pp.205-228). Sussex: Psychology Press.

O'Sullivan, R. G. (2012). Collaborative evaluation within a framework of stakeholder-oriented evaluation approaches. *Evaluation and program planning*, *35*(4), 518-522.

Palumbo, D. J., & Nachmias, D. (1983). The preconditions for successful evaluation: Is there an ideal paradigm? *Policy Sciences*, *16*(1), 67-79.

Patel, L. (2008). Getting it right and wrong: An overview of a decade of post-apartheid social welfare. *Practice*, *20*(2), 71-81.

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. D*ialogues Clin Neurosci*, *13*(2), 217-224.

Patton, M. Q. (2008). *Utilization-Focused Evaluation*. Thousand Oaks, CA: Sage Publications Ltd.

Patton, M. Q. (2013). *Nomination justification statement.* Retrieved from http://www.cgu.edu/PDFFiles/sbos/donaldson/Patton%20Nomination%20of%20Donaldson%20Lasarsfeld%20Oct%202013.pdf

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: SAGE Publications.

Pawson, R., & Tilley, N. (2004). How to construct realistic data utilizing stakeholders' knowledge. In *Realistic evaluation* (pp. 153–182). Thousand Oaks, CA: SAGE Publications

Peersman, G., Guijt, I., & Pasanen, T. (2015). *Evaluability assessment for impact evaluation: Guidance, checklists and decision support*. London, England: A Methods Lab.

Petrucci, C. J. (2009). A primer for social worker researchers on how to conduct a multinomial logistic regression. *Journal of Social Service Research*, *35*(2), 193–205.

Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *Academy of Management Review*, *18*(4), 599-620.

Picciotto, R. (2003). International trends and development evaluation: the need for ideas. *American Journal of Evaluation*, *24*(2), 227–234.

Podems, D., Goldman, I., & Jacob, C. (2014). Evaluator competencies: The South African government experience. *Canadian Journal of Program Evaluation, 28*(3), 71-85.

Pusha, S., Gudi, R., & Noronha, S. (2009). Polar classification with correspondence analysis for fault isolation. *Journal of Process Control*, *19*(4), 656-663.

Rabie, B. (2010). An exploration of South Africa's framework for public sector monitoring and evaluation: Lessons from international best practice. *Administratio Publica, 18*(1), 139–160. Retrieved from http://journal.assadpam.net/index.php?journal=assadpam

Ramlo, S. (2005). An application of Q methodology: Determining college faculty perspectives and consensus regarding the creation of a school of technology. *Journal of Research in Education, 15*(1), 52-69.

Ramlo, S., & Newman, I. (2011). Classifying individuals using Q methodology and Q factor analysis: Applications of two mixed methodologies for program evaluation. *J Res Educ*, *21*, 20–31. Retrieved from http://www.academia.edu/

Ramlo, S., & Newman, I. (2011). Classifying individuals using Q methodology and Q factor analysis: applications of two mixed methodologies for program evaluation. *J Res Educ, 21*, 20-31.

Rasp, A. (1981). Book reviews: Planning useful evaluations: Evaluability assessment. *Educational Evaluation and Policy Analysis, 3*(3), 104–106.

Reber, B. H., Kaufman, S. E., & Cropp, F. (2000). Assessing Q-Assessor: A validation study of computer-based Q sorts versus paper sorts. *Operant Subjectivity*, *23*(4), 192–209. Retrieved from http://q-assessor.com/versions/3000

Reichardt, C. S. (2011). Evaluating methods for estimating program effects. *American Journal of Evaluation, 33*(2), 246–272.

Rist, R. C. (1989). Management accountability: The signals sent by auditing and evaluation. *Journal of Public Policy, 9*(3), 355–369. Retrieved from http://www.jstor.org/stable/pdf/4007444.pdf

Rist, R. C. (1990). The organization and function of evaluation in the United States: A federal overview. In R. C. Rist (Ed.), *Program evaluation and the management of government: Patterns and prospects across eight nations* (pp. 71–94). New Brunswick, NJ: Transaction Publishers.

Rist, R. C., & Paliokas, K. L. (2002). The rise and fall (and rise again?) of the evaluation function in the US government. In J. E. Furubo, R. C. Rist, & R. Sandahl (Eds.), *International atlas of evaluation* (pp. 225–248). London: Transaction Publishers.

Rodriguez-Campos, L. (2011). Stakeholder involvement in evaluation: three decades of the American journal of evaluation. *Journal of Multidisciplinary Evaluation*, *8*(17), 57-79.

Roelofse-Campbell, Z. (2006). Post-apartheid South Africa and Brazil: A strategic partnership. *UNISA Latin American Report*, *22*(1), 92-108.

Rog, D. J. (1997, March). When NOT to do an outcome evaluation: Assessing the evaluability of a program. In R. E. Stake (Chair), *Grounds for turning down a handsome evaluation contract.* Symposium conducted at the meeting *of* the American Educational Research Association, Chicago, IL.

Rog, D. J. (2012). When background becomes foreground: Toward context-sensitive evaluation practice. *New Directions for Evaluation, 2012*(135), 25–40.

Rogers, P. J., Petrosino, A., Huebner, T. A., & Hacsi, T. A. (2000). Program theory evaluation: Practice, promise, and problems. *New Directions for Evaluation*, *2000*(87), 5–13.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: SAGE Publications.

Rourke, J. T. (1978). The GAO: An evolving role. *Public Administration Review, 38*(5), 453–457. Retrieved from http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1540-6210

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, *25*(1), 127-141.

Rutman, L. (1980). *Planning useful evaluations: Evaluability assessment*. Beverly Hills, CA: SAGE Publications.

SAMEA (2016). *About us*. Retrieved from http://www.samea.org.za/samea-2.phtml

Schmidt, R. E., Scanlon, J. W., & Bell, J. B. (1979). *Evaluability assessment: Making public programs work better* (Monograph No. 14). Rockville, MD: Project SHARE.

Schön, D. A. (1983). The reflective practitioner: How professionals think in action (Vol. 5126). Basic Books, Inc.

Schwandt, T. A. (2003). Back to the rough ground! Beyond theory to practice in evaluation. *Evaluation*, *9*(3), 353–364.

Schwandt, T. A. (2005). The centrality of practice to evaluation. *American Journal of Evaluation*, *26*(1), 95–105.

Schwandt, T. A. (2007). Expanding the conversation on evaluation ethics. *Evaluation and Program Planning*, *30*(4), 400–403.

Schwandt, T. A. (2008). Educating for intelligent belief in evaluation. *American Journal of Evaluation*, *29*(2), 139–150. doi:10.1177/1098214008316889

Schwandt, T., & Dahler-Larsen, P. (2006). When evaluation meets the 'rough ground' in communities. *Evaluation, 12*(4), 496–505.

Seo, S. H. (2008). *A study on democratic transition in South Africa: Democracy through compromise and institutional choice.* (Doctoral dissertation, University of South Africa). Retrieved from http://uir.unisa.ac.za/bitstream/handle/10500/3401/thesis_seo_s.pdf?sequence=1

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston, MA: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: SAGE Publications.

Shadish, W. R., & Epstein, R. (1987). Patterns of program evaluation practice among members of the Evaluation Research Society and Evaluation Network. *Evaluation Review*, *11*(5), 555-590.

Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *American Journal of Evaluation*, *18*(3), 195–208. Retrieved from http://aje.sagepub.com/

Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Boston, MA: Kluwer Academic.

Smith, M. F. (1990). Evaluability assessment: Reflections on the process. *Evaluation and Program Planning, 13*(4), 359–364.

Smith, N. L. (1981). Evaluability assessment: A retrospective illustration and review. *Educational Evaluation and Policy Analysis, 3*(1), 77–82. Retrieved from http://www.jstor.org/

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *American Journal of Evaluation*, *14*(3), 237–242.

Smith, N. L. (1998). Professional reasons for declining an evaluation contract. *American Journal of Evaluation, 19(2)*, 177-190.

South African Social Security Agency (2014). *A statistical summary of social grants in South Africa Fact sheet* (Issue no 2 of 2014 – 31 February 2014). Retrieved from www.sassa.gov.za/index.php/knowledge-centre/statistical-reports?download=207:statistical-report-2-of-2014 south african social grant stat

SPSS Inc. (2006). *Advanced statistical analysis using SPSS.* Chicago, IL: SPSS Inc.

Stainton Rogers, R. (1995). Q methodology. In I. A. Smith, R. Harre, & L. Van Langenhove (Eds.), *Rethinking methods in psychology* (pp. 178-192). Thousand Oaks, CA: Sage Publications.

Stake, R. E. (1976). A theoretical statement of responsive evaluation. *Studies in Educational Evaluation, 2*(1), 19-22.

Stake, R. E. (1990). Situational context as influence on evaluation design and use. *Studies in Educational Evaluation, 16*(2), 231-246.

Stame, N. (2003). Evaluation and the policy context: The European experience. *Evaluation Journal of Australasia, 3*(2), 36–43. Retrieved from http://www.aes.asn.au/evaluation-journal-of-australasia.html

Stephenson, W. (1935). Correlating persons instead of tests. *Journal of Personality, 4*(1), 17-24.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation, 26*(1), 43–59.

Strosberg, M. A., & Wholey, J. S. (1983). Evaluability assessment: From theory to practice in the department of health and human services. *Public Administration Review, 43*(1), 66–71. Retrieved from http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1540-6210

Stufflebeam, D. L. (1971). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education, 5*, 19-25.

Talbot, C. (2010). *Performance in government: The evolving system of performance and evaluation measurement, monitoring, and management in the United Kingdom.* Washington, DC: The World Bank.

Tanner, M. (2012). *The American welfare state. How we spend nearly $1 trillion a year fighting poverty—and fail* (Policy Analysis No. 694). Retrieved from http://heartland.org/sites/default/files/pa694.pdf

Taut, S., & Brauns, D. (2003). Resistance to evaluation a psychological perspective. *Evaluation, 9*(3), 247–264. doi:10.1177/13563890030093002

Taylor, N., Muller, J., & Vinjevold, P. (2003). *Getting schools working: Research and systemic school reform in South Africa*. South Africa: Pearson Education.

The Commonwealth. (2016). United Kingdom: Constitution and politics. Retrieved from http://thecommonwealth.org/our-member-countries/united-kingdom/constitution-politics

The International Bank for Reconstruction and Development / the World Bank (2004). Brazil: Forging a strategic partnership for results. An OED evaluation of World Bank assistance. Retrieved from http://lnweb90.worldbank.org/oed/oeddoclib.nsf/DocUNIDViewForJavaSearch/817C4FB038C4427E85256E2A0074AE96/$file/brazil_cae.pdf

The World Bank (2015). World Bank country and lending groups. Retrieved https://datahelpdesk.worldbank.org/knowledgebase/articles/906519

The World Bank (2016a). *World Bank open data*. Retrieved from http://data.worldbank.org/country/brazil.

The World Bank (2016b). *Countries*. Retrieved from http://www.worldbank.org/en/country/brazil/overview

The World Bank (2016c). *Projects and operations*. Retrieved from http://www.worldbank.org/projects/P087713/br-bolsa-familia-1st-apl?lang=en

The World Bank (2016d). *World Bank open data*. Retrieved from http://data.worldbank.org/country/southafrical.

The World Bank (2016e). *Countries*. Retrieved from http://www.worldbank.org/en/country/southafrica/overview

Thomas, D. B., & Baas, L. R. (1992). The issue of generalization in Q methodology: Reliable schematics revisited. *Operant Subjectivity, 16*(1), 18-36.

Thompson, B. (1998, May). *Using Q-technique factor analysis in education program evaluations or research: An introductory primer*. Paper presented at the Annual Conference on Research Innovations in Early Intervention, Charleston, SC.

Thompson, B., & Miller, L. A. (1983, May). *Differences and Similarities in Administrators' and Evaluators' Perceptions of Evaluation*. Paper presented at the 67th Annual Meeting of the American Educational Research Association, Montreal, Quebec.

Thurston, W. E., & Potvin, L. (2003). Evaluability assessment: A tool for incorporating evaluation in social change programmes. *Evaluation, 9*(4), 453–469.

Thurston, W. E., Graham, J., & Hatfield, J. (2003). Evaluability assessment. A catalyst for program change and improvement. *Evaluation and the Health Professions, 26*(2), 206–221. doi:10.1177/0163278703252264

Tilley, N. (2000, September). *Realistic evaluation: An overview*. Paper presented at the meeting of the Danish Evaluation Society, Denmark.

Toulemonde, J. (1995). The emergence of an evaluation profession in European countries: The case of structural policies. *Knowledge and Policy, 8*(3), 43–54.

Tourmen, C. (2009). Evaluators' decision making the relationship between theory, practice, and experience. *American Journal of Evaluation, 30*(1), 7–30.

Trevisan, M. S. (2004). Practical training in evaluation: A review of the literature. *American Journal of Evaluation*, *25*(2), 255-272.

Trevisan, M. S. (2007). Evaluability assessment from 1986 to 2006. *American Journal of Evaluation, 28*(3), 290–303.

Trevisan, M. S., & Huang, Y. M. (2003). Evaluability assessment: A primer. *Practical Assessment, Research & Evaluation, 8*(20), 2–9. Retrieved from http://pareonline.net/

Trevisan, M. S., & Walser, T. M. (2014). Evaluability assessment: Improving evaluation quality and use. Sage Publications, Inc.

Tubergen, N. V., & Olins, R. A. (1978). Mail vs. personal interview administration for Q sorts: A comparative study. *Operant subjectivity, 2*(2), 51-59.

U.S. Government Accountability Office. (2011). *Program evaluation: Experienced agencies follow a similar model for prioritizing research* (GAO 11-176). Retrieved from http://www.gao.gov/new.items/d11176.pdf

United Nations Development Fund for Women (UNIFEM). (2009). Guidance note on carrying out an evaluability assessment. Retrieved fromhttp://erc.undp.org/unwomen/resources/guidance/Guidance%20Note%20-%20Carrying%20out%20an%20Evaluability%20Assessment.pdf

University of Bern, Centre for University Continuing Education (2012). *European university-based study programmes in evaluation: Sixteen profiles*. Retrieved from http://europeanevaluation.org/sites/default/files/16_profiles_November%202012.pdf

US Agency, International Development. (2008). *Planning for cost effective evaluation with evaluability assessment* (Impact Assessment Primer Series Publication No. 6). Retrieved from http://pdf.usaid.gov/pdf_docs/PNADN200.pdf

UK Evaluation Society. (2016). *About us*. Retrieved from https://www.evaluation.org.uk/index.php/about-us

van Exel, J., & de Graaf, G. (2005). *Q methodology: A sneak preview*. Retrieved from https://www.researchgate.net/publication/228574836_Q_Methodology_A_Sneak_Preview

Van Voorhis, P., & Brown, K. (1996). *Evaluability assessment: A tool for program development in corrections* [Monograph]. Retrieved from https://s3.amazonaws.com/static.nicic.gov/Library/014292.pdf

Vergnaud, G., Pastré, P., & Mayen, P. (2006). La didactique professionnelle (note de synthèse). *Revue française de pédagogie*, (154), 145-198.

Victora, C. G., Habicht, J. P., & Bryce, J. (2004). Evidence-based public health: moving beyond randomized trials. *American journal of public health, 94*(3), 400-405.

Wargo, M. J. (1995). The impact of federal government reinvention on federal evaluation activity. *American Journal of Evaluation*, *16*(3), 227–237.

Watts, B. R., & Washington, H. M. (2016). Adaptation and Use of a Five-Task Model for Evaluability Assessment. *Journal of MultiDisciplinary Evaluation, 12*(27), 67-78.

Watts, S., & Stenner, P. (2005). Doing Q methodology: theory, method and interpretation. *Qualitative Research in Psychology, 2*(1), 67-91.

Watts, S., & Stenner, P. (2012). *Doing Q methodological research: Theory, method & interpretation*. Thousand Oaks, CA: Sage Publications Ltd.

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53–85. Retrieved from http://www.annualreviews.org/journal/psych

Weiss, C. H. (1998). Purposes of evaluation. In *Evaluation: Methods for studying programs and policies* (2nd ed., pp. 18–42). Upper Saddle River: Prentice Hall.

White, H., Sabarwal, S., & de Hoop, T. (2014). *Methodological briefs: Randomized controlled trials (RCTs)*. United Nations Children's Fund (UNICEF) Office of Research. Retrieved from https://www.unicef-irc.org/publications/pdf/brief_7_randomized_controlled_trials_eng.pdf

White, J. L. (2013). Logistic regression model effectiveness: proportional chance criteria and proportional reduction in error. *Journal of Contemporary Research in Education*, *2*(1), 4-10.

Wholey, J. S. (1979). *Evaluation: Promise and performance*. Washington, DC: Urban Institute.

Wholey, J. S. (1994). *Assessing the feasibility and likely usefulness of evaluation*. San Francisco, CA: Jossey-Bass Publisher.

Wholey, J. S. (2004). Evaluability assessment. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.) *Handbook of practical program evaluation* (2nd ed., pp. 33–62). San Francisco, CA: Josey-Bass.

Wholey, J. S. (2010). Exploratory evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.) *Handbook of practical program evaluation*, (3rd ed., pp. 81–99). San Francisco, CA: Josey-Bass.

Wholey, J. S., Nay, J. N., Scanlon, J. W., & Schmidt, R. E. (1975). Evaluation: When is it really needed? *Evaluation, 2*(2), 89–93. Retrieved from https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=48470

Widmer, T. (2004). The development and status of evaluation standards in Western Europe. *New Directions for Evaluation, 2004*(104), 31–42.

Worden, N. (2011). *The making of modern South Africa: conquest, apartheid, democracy*. Sussex: John Wiley & Sons Ltd.

Worthen, B. (1995). The unvarnished truth about logic-in-use versus reconstructed logic in educational inquiry. *Evaluation Practice*, *16*(2), 165–178.

Worthen, B. R. (1994). Is evaluation a mature profession that warrants the preparation of evaluation professionals? *New Directions for Program Evaluation, 1994*(62), 3–15.

Yelland, P. M. (2010). An introduction to correspondence analysis. *The Mathematica Journal*, *12*, 1-23.

Youker, B. W. (2013). Goal-free evaluation: A potential model for the evaluation of social work programs. *Social Work Research*, *37*(4), 432–438.

Appendix A

Evaluability Criteria Derived from the Literature

Table A1

Summary of Evaluability Criteria Derived from Literature

Author	Evaluability criteria
Chen (2005)	Well-defined programme goals and objectives
	Realistic and plausible goals and objectives
	Sufficient and easily obtainable performance data
	Stakeholders agreement on how evaluation will be used
Newcomer et al. (2010)	Potential influence of evaluation on decision making
	Timing / Feasibility
	Significance
	Perceived programme performance
	Programme life cycle
Wholey (1979; 2010)	Clearly defined, measurable and agreed upon objectives
	Explicit and plausible programme theory
	Programme activities implemented as planned
	Specified and measurable indicators of programme implementation and performance
	Clearly identified information needs
	Evaluation data are obtainable
Horst et al. (1979)	Clear definition (problem to be addressed, intervention, desired outcomes)
	Clear logic (well-articulated programme theory)
	Programme management (adequate motivation, ability and authority to facilitate the evaluation process and act on evaluation findings)
UNIFEM Evaluation Unit (2008)	Clear programme design (clearly defined problem, target beneficiaries, and desired outcomes/impact) and plausible theory of change
	Availability of data (capacity to provide data, monitoring system, accessibility of data, costs of data collection and analysis)
	Conduciveness of context (evaluation capacity and expertise, evaluation resources)

Table A1 cont.

Summary of Evaluability Criteria Derived from Literature

Author	Evaluability Criteria
Smith (1981)	Need for accountability
	Amount of resources invested in programme design and implementation
	Potential social benefits of the programme and anticipated magnitude of these benefit
	Level of public interest in the programme
	Relevance of evaluative information to future policy formulation or programme decision
Juvenile Justice Evaluation Centre (2003)	Explicit programme design
	Plausible programme theory
	Measurable outcomes
	Programme serves the population for whom it was designed
	Programme has the resources discussed in the programme design
	Programme activities implemented as designed
	Programme have the capacity to provide data for the evaluation
Rutman (1980)	Programme (or its components) clearly defined and capable of being implemented in a prescribed manner
	Goals and effects clearly specified
	Programme can realistically achieve the specified goals or produce the anticipated effects?
	Feasibility of implementing desired methodology to meet the purposes of the evaluation:
	Limited restrictions placed on the evaluation by various constraints- financial, political, legal, ethical and administrative
Schmidt, Scalon,& Bell (1980)	Programme description complete
	Programme description acceptable to policymakers
	The expectations of the programme are plausible
	The evidence required by management can be reliably produced
	The evidence required by management is feasible to collect
	Management's intended use of the information can realistically be expected to affect performance

Table A1 cont.

Summary of Evaluability Criteria Derived from Literature

Author	Evaluability Criteria
Stenberg (1983)	Programme objectives are well defined (i.e., those in charge of the programme have agreed on a set of realistic, measurable objectives and programme performance indicators in terms of which programme is to be held accountable and managed)
	Programme objectives are plausible
	Intended use of information is well-defined (i.e., those in charge of the programme have agreed on how program performance information will be used to achieve improved programme performance)
Taut & Brauns (2003); Leviton (2010); Jung & Shuberg (1983)	Stakeholders willingness to engage in evaluation process
	Stakeholders willingness to make changes on basis of evaluation findings
	Consensus among stakeholders and evaluator (performance criteria, evaluation questions, design, cost and timeline of evaluation)
Davies (2013)	*Programme design*
(Synthesis based on an examination of guidance documents produced by: UNIFEM, USAID,ILO,IADB; AusAID, IDRC, EBRD, and NDC)	Clearly identified long-term impact and outcomes
	Clearly defined and plausible causal links
	Project objective clearly relevant to the needs of the target group
	Intended beneficiary group clearly identified
	Valid and reliable indicators of expected event
	Assumptions about the roles of other actors outside the project made explicit
	Possible to identify which linkages in the causal chain will be most critical to the success of the project
	Consistency in the way the Theory of Change is described across various project multiple documents
	Visibility of stakeholder views

Table A1 cont.

Summary of Evaluability Criteria Derived from Literature

Author	Evaluability Criteria
Davies (2013)	*Information availability*
	Complete set of documents available
	Availability of baseline data
	Availability of control data
	Data collected for all indicators with sufficient frequency
	Critical data available (e.g., process data; evaluation reports)
	Existing M&E systems have the capacity to deliver
	Institutional context
	Accessibility to and availability of stakeholders
	Resources available to do the evaluation
	Correct evaluation timing
	Primary users been clearly identified
	Realistic evaluation questions
	Design feasible
	Manageable ethical issues
	Manageable risks

Appendix B

Ethical Clearance

Courier: Room 2.21 Leslie Commerce Building Upper Campus University of Cape Town
Post: University of Cape Town Private Bag Rondebosch 7701
Email: Irwin.brown@uct.ac.za
Telephone: +27 21 650-2311
Fax No.: +27 21 689-7570

March 17, 2014
Adiilah Boodhoo
Management Studies

Dear Researcher

Project title:

Evaluator characteristics and programme evaluability decisions

This letter serves to confirm that this project as described in your submitted protocol has been approved. Please note that if you make any substantial change in your research procedure that could affect the experiences of the participants, you must submit a revised protocol to the Committee for approval.

Regards,
Harold Kincaid
Professor Harold Kincaid
Commerce Faculty Ethics in Research Committee

Appendix C

QSort Task

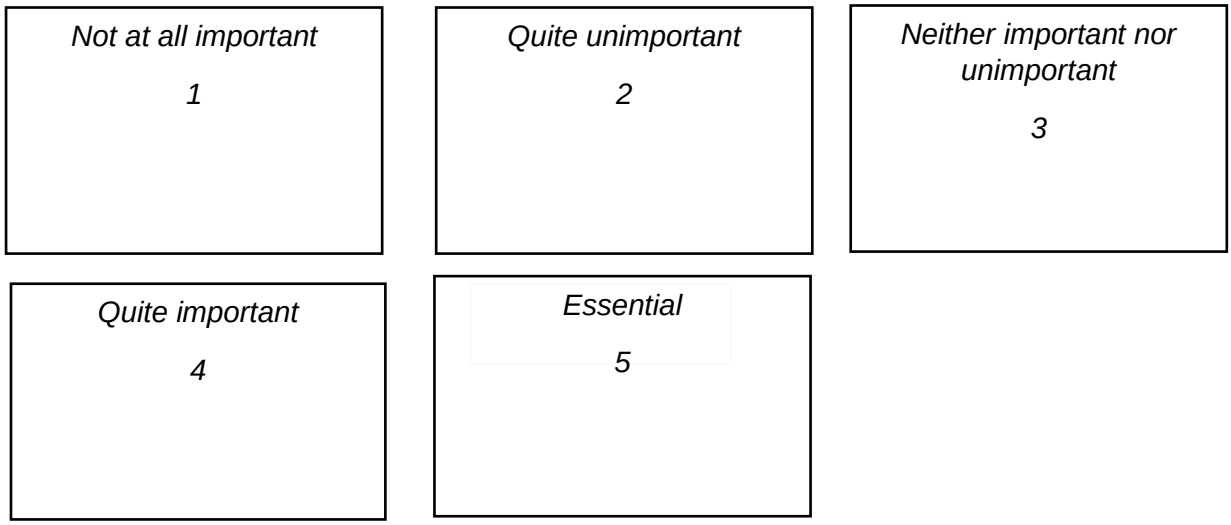
Instructions:

Step 1: On the left of the screen you will find 19 statements. These statements represent the criteria that you might consider when deciding how evaluable a programme is. Read each statement carefully.

Statements

1. Programme goals are clearly specified
2. Programme outcomes are realistic
3. Programme outcomes are measurable
4. Stakeholders agree on programme goals
5. Programme data are adequate
6. Programme data are reliable
7. Programme data are easily accessible
8. Programme theory is explicitly stated
9. Programme theory is plausible
10. The manner in which the programme is delivered is clearly defined
11. Target beneficiaries are clearly defined
12. Programme is implemented as intended.
13. Stakeholders are willing to collaborate with the evaluator
14. Stakeholders have authority to act on evaluation findings
15. Purpose for which evaluation results will be used is clear
16. Budget is adequate for the evaluation
17. Timeframe is adequate to complete the evaluation
18. Type of evaluation required (process, outcome or impact) is feasible
19. Required evaluation methodology is feasible

Step 2: On the right of the screen you will find five boxes labelled *Not at all important, Quite unimportant, Neither important nor unimportant, Quite important, Essential.* Please drag each statement into the box that you think best represents how important the criterion is when deciding how evaluable a programme is.



Which of the criteria placed in box 5 (Essential) do you prioritise the most when assessing the evaluability of a programme?

If you placed two or more statements in box 5, indicate which ones would appear first, second or third* on your high priority list. You may simply insert the number attached to the relevant criteria in the textboxes below.

First on high priority list:

Second on high priority list:

Third on high priority list:

Which of the criteria placed in box 1 (Not at all important) do you prioritise the least when assessing the evaluability of a programme?

If you placed two or more statements in box 1, indicate which ones would appear first, second or third* on your low priority list. You may simply insert the number attached to the relevant criteria in the textboxes below.

First on low priority list:

Second on low priority list:

Third on low priority list:

Appendix D

Evaluability Scenarios

Background

We would like you to imagine that…

A funding agency wants to commission an outcome evaluation of an educational support programme for high school students. The programme is a one year, post-high school intervention for students who did not gain entry into a tertiary institution. The programme offers a variety of academic activities. The client would like you to compare the performance scores of the beneficiaries with that of a comparison group who did not receive the programme. You have had an initial meeting with the programme staff and need to decide whether or not you will accept the evaluation contract.

Instructions:

You will be presented with three short scenarios of the programme to be evaluated. Each scenario will be presented in a different colour (either blue, green or purple). Please read each scenario carefully answer the three questions that follow. There are no predetermined right or wrong responses to these questions. We request you to approach this exercise in the same manner that you would have if this were a real life situation.

Each scenario will be presented only once. It might be useful to take down notes of what you think is important for your decision making as you read the scenarios. Please click on the 'NEXT' button to proceed.

Scenario 1

What the programme does:

Prepare selected beneficiaries to re-write their secondary school leaving certificate and apply for tertiary education. The aim is to improve the access of socio-economically disadvantaged students into tertiary education.

Who gets the programme:

Students who have finished high school in the previous year but who did not meet the minimum performance requirements for entry into a tertiary institution or their degree of choice. These students come from socio-economically disadvantaged communities.

What happens on the programme:

The programme offers intensive academic tutoring, personal mentoring, and assistance with tertiary applications.

What change will the programme bring about:

There is consensus among programme stakeholders that the expected outcomes of the programme are: improved student performance on examination re-write and improved student access to preferred tertiary institutions and field of study.

What is available to the evaluator:

You will have unrestricted access to (1) the programme's electronic monitoring system, which includes verified pre-programme and post-programme school performance scores of beneficiaries; (2) programme records, including the logic framework and (3) external evaluation reports confirming implementation fidelity and the plausibility of the programme theory. You will have to find and collect data from a suitable matched comparison group.

How much money and time is available:

There is a very tight budget for the evaluation and the evaluation report has to be completed within 10 working days.

Characteristics of programme staff:

Your initial client meeting was characterised by a high level of stakeholder animosity and resistance to engage in the evaluation process. Programme staff were not clear on how the results of the evaluation will be used. They do not have the authority to implement the findings of external evaluations commissioned by the funding agency.

Scenario 2

What the programme does:

Different programme goals are specified in different programme documents.

Who gets the programme:

The criteria used to select beneficiaries into the programme vary across different programme intakes.

What happens on the programme:

Over the years, a number of changes have been made to programme activities.

What change will the programme bring about:

There is little consensus among programme stakeholders regarding the expected outcomes of the programme. Some stakeholders focus on student's socio emotional welfare, while others regard improved employment prospects after completion of tertiary studies as the main outcome.

What is available to the evaluator:

The programme does not have an electronic monitoring system. Only pre-programme data for programme recipients have been collected on paper forms by the programme staff. You will have access to these paper records upon request. There are no records of the programme's logical framework, implementation fidelity or the plausibility of the programme theory. You will have to find and collect data from a suitable matched comparison group.

How much money and time:

There is a very tight for the evaluation and the evaluation report has to be completed within 10 working days.

Characteristics of programme staff:

Your initial client meeting was characterised by a high level of enthusiasm and stakeholder willingness to engage in the evaluation process. Programme staff are clear on how they will use the evaluation results. They have the authority to implement the findings of external evaluations commissioned by the funding agency.

Scenario 3

What the programme does:

Different programme goals are specified in different programme documents.

Who gets the programme:

The criteria used to select beneficiaries into the programme vary across different programme intakes.

What happens on the programme:

Over the years, a number of changes have been made to programme activities.

What change will the programme bring about:

There is little consensus among programme stakeholders regarding the expected outcomes of the programme. Some stakeholders focus on student's socio emotional welfare, while others regard improved employment prospects after completion of tertiary studies as the main outcome.

What is available to the evaluator:

The programme does not have an electronic monitoring system. Only pre-programme data for programme recipients have been collected on paper forms by the programme staff. You will have access to these paper records upon request. You will also have access to a matched comparison group. There are no records of the programme's logical framework, implementation fidelity or the plausibility of the programme theory.

How much money and time is available:

There is a generous budget for the evaluation and the evaluation report has to be completed within 20 working days.

Characteristics of programme staff:

You initial client meeting was characterised by a high level of stakeholder animosity and resistance to engage in the evaluation process. Programme staff were not clear on how the results of the evaluation will be used. They do not have the authority to implement the findings of external evaluations commissioned by the funding agency.

---

1. **In your opinion, to what extent is this programme evaluable?**

Evaluable with a lot of difficulty — Evaluable with minimal difficulty

1 2 3 4 5 6 7 8 9 10

Participants with a score of 1-5 on item 1 were required to respond to the following three items:

2. **You indicated that this programme was evaluable with a lot of difficulty. Please provide three reasons to explain your assessment. State these reasons in order of importance.**

Somewhat Important [Text box]

Very Important [Text box]

Essential [Text box]

3. **How likely are you to evaluate this programme?**

Very unlikely — Very likely

1 2 3 4 5 6 7 8 9 10

4. **If you have to evaluate this programme, what actions would you take before you conduct the evaluation?**

[Text box]

[Text box]

[Text box]

Participants with a score of 6-10 on item 1 were required to respond to the following item two items:

2. **You indicated that this programme was evaluable with minimal difficulty. Please provide three reasons to explain your assessment. State these reasons in order of importance.**

Somewhat Important	[Text box]
Very Important	[Text box]
Essential	[Text box]

3. How likely are you to evaluate this programme?

Very unlikely Very likely

1 2 3 4 5 6 7 8 9 10

Appendix E

Evaluator Characteristics

Table E1

Evaluator Characteristics

	Brazil		SA		UK		USA	
	n	%	*N*	%	*n*	%	*n*	%
Current Involvement in Evaluation								
Design evaluations	28	30.8	24	53.5	19	63.3	72	76.6
Conduct evaluations	48	52.7	29	64.4	22	73.3	71	75.5
Lead team of evaluators	11	12.1	13	28.9	15	50.0	40	42.6
Employed in evaluation job	10	11.0	22	48.9	14	46.7	64	68.1
Academic interest in evaluation	58	63.7	19	42.2	8	26.7	40	42.6
Publish on evaluation	25	27.5	7	15.6	9	30.0	35	37.2
Other	7	7.7	2	4.4	1	3.3	6	6.4
Employment Setting								
University	24	26.4	5	11.1	4	13.3	19	20.2
Public sector	37	40.7	6	13.3	3	10.0	16	17.0

Table E1 cont.

Evaluator Characteristics

	Brazil		SA		UK		USA	
	n	%	*N*	%	*n*	%	*n*	%
Employment Setting								
Private evaluation consultancy firm	2	2.2	7	15.6	8	26.7	15	16.0
Own evaluation consultancy	4	4.4	6	13.3	4	13.3	13	13.8
NGO/NPO	6	6.6	10	22.2	2	6.7	14	14.9
Other	4	4.4	4	8.9	3	10.0	9	9.6
Missing data	12	13.2	7	15.6	4	13.3	8	8.5
Highest Academic Qualification								
Undergraduate	7	7.7	1	2.2	4	13.3	4	4.3
Postgraduate diploma	15	16.5	26	57.8	0	0.0	1	1.1
Master’s degree	32	35.2	11	24.4	16	53.3	33	35.1
PhD	23	25.3	38	84.4	4	13.3	47	50.0
Other	2	2.2	11	24.4	2	6.7	1	1.1
Missing data	12	13.2	7	15.6	4	13.3	8	8.5

Table E1 cont.

Evaluator Characteristics

	Brazil		SA		UK		USA	
	n	%	*N*	%	*n*	%	*n*	%
Academic Discipline								
Psychology	6	6.6	8	17.8	0	0.0	19	20.2
Education	3	3.3	2	4.4	1	3.3	12	12.8
Sociology/Social work	1	1.1	3	6.6	1	3.3	6	6.4
Evaluation	13	14.3	9	20.0	1	3.3	9	8.5
Public Health	4	4.4	2	4.4	1	3.3	7	7.4
Public Policy and Administration	10	11.0	1	2.2	1	3.3	9	9.6
Other	35	38.5	11	24.4	21	70.0	24	25.5
Missing data	19	20.9	9	20.0	4	13.3	8	8.5
Type of Training in Evaluation								
Self-educated	17	18.7	15	33.3	19	63.3	28	29.8
Short course certificate in evaluation	27	29.7	7	15.6	3	10.0	11	11.7

Table E1 cont.

Evaluator Characteristics

	Brazil		SA		UK		USA	
	n	%	*N*	%	*n*	%	*n*	%
Type of Training in Evaluation								
Postgraduate diploma	9	9.9	3	6.7	0	0.0	1	1.1
Master's degree	15	16.5	11	24.4	1	3.3	17	18.1
PhD	11	12.1	2	4.4	3	10.0	29	30.9
Missing data	12	13.2	7	15.6	4	13.3	8	8.5
Years Conducting Evaluation								
Less than a year	12	13.2	3	6.7	0	0.0	1	1.1
1 to 5 years	37	40.7	15	33.3	5	16.7	27	28.7
6 to 10 years	13	14.3	8	17.8	11	36.7	19	20.2
11 to 15 years	11	12.1	7	15.6	2	6.7	12	12.8
More than 15 years	5	5.5	5	11.1	8	26.7	27	28.7
Missing data	13	14.3	7	15.6	4	13.3	8	8.5

Table E1 cont.

Evaluator Characteristics

	Brazil		SA		UK		USA	
	n	%	*n*	%	*n*	%	*n*	%
Time Since Last Evaluation								
Currently working on an evaluation	0	0.0	29	63.0	21	70.0	67	71.3
Less than a month ago	40	44.0	4	9.7	0	0.0	3	3.2
1 to 6 months ago	4	4.4	1	2.2	3	10.0	8	8.5
7 to 12 months ago	12	13.2	3	6.5	1	3.3	3	3.2
More than a year ago	9	9.9	2	4.3	1	3.3	5	5.3
Missing data	26	28.6	7	15.2	4	13.3	8	8.5
Number of Evaluations Completed in the Last 5 years								
1-5	49	53.8	17	37.8	7	23.3	28	29.8
6-10	6	6.6	11	24.4	5	16.7	19	20.2
11-15	5	5.5	5	11.1	4	13.3	15	16.0
16-20	3	3.3	2	4.4	2	6.7	8	8.5
21-25	2	2.2	1	2.2	4	13.3	4	4.3

Table E1 cont.

Evaluator Characteristics

	Brazil		SA		UK		USA	
	n	%	*n*	%	*n*	%	*n*	%
Number of Evaluations Completed in the Last 5 years								
Over 25	0	0.0	1	2.2	3	10.0	6	6.4
Missing data	26	28.6	8	17.8	5	16.7	14	14.9
Practice Context								
Developing countries	71	78.0	36	78.3	10	33.3	7	7.4
Developed countries	2	2.2	1	2.2	12	40.0	75	79.8
Both	5	5.5	2	4.3	4	13.3	4	4.3
Missing data	13	14.3	7	15.2	4	13.3	8	8.5

Appendix F

Evaluator Profile Items

Just a few questions that will allow us to describe the overall characteristics of evaluators who participated in this study. This data is critical to the study. Please click on the relevant response option to indicate your answer.

1. **What is your current involvement in evaluation? (You may select more than one option here).**

   1 – I design evaluations
   2 – I conduct evaluations
   3 – I lead a team of evaluators
   4 – I am employed in an evaluation job
   5 – I have an academic interest in evaluation
   6 – I publish on evaluation
   7 – Other
   8 – I do not work in evaluation

2. **In which setting are you currently employed?**

   1 – In a university
   2 – In the public sector
   3 – I work for a private evaluation/research consultancy firm
   4 – I have my own evaluation practice
   5 – I work for an international donor agency
   6 – I work for an non-governmental/non-profit organization
   7 – Other Please specify: [Text box]

3. **What is your highest academic qualification?**

   1 – Undergraduate degree
   2 – Post graduate diploma
   3 – Master's degree
   4 – PhD
   5 – Other Please specify: [Text box]

4. **In which discipline did you obtain your highest qualification?** [Text box]

5. What type of training/education have you had in programme evaluation?

1 – Self-educated
2 – Short course certificate in evaluation
3 – Post graduate diploma
4 – Master's degree
5 – PhD

6. For how long have you been conducting evaluations?

1 – Less than a year
2 – 1 to 5 years
3 – 6 to 10 years
4 – 11 to 15 years
5 – More than 15 years
6 – I do not conduct evaluations

7. When did you conduct your last evaluation?

1 – I am currently working on an evaluation
2 – Less than a month ago
3 – 1 to 6 months ago
4 – 7 to 12 months ago
5 – More than a year ago
6 – I do not conduct evaluations

8. How many evaluations have conducted in the last 5 years (if any)? [Text box]

9. **Indicate your level of experience in conducting the following:**

	Not at all experienced	Slightly experienced	Moderately experienced	Highly experienced
	1	*2*	*3*	*4*
Evaluation readiness assessments				
Needs assessments				
Implementation/process evaluations				
Outcome evaluations				
Impact evaluations				
Summative evaluations				
Formative evaluations				
Meta-analysis of evaluations				

10. Where do you mostly do evaluation work?

1 – In developing countries

2 – In developed countries

3 – In both developing and developed countries

11. Your name was selected from the UKES/SAMEA/AEA/BMEN membership list. Do you belong to any other evaluation associations?

1 – Yes Please specify: [Text box]

2 – No

Appendix G

Missing Data

Table G1

Valid and Missing Data Per Country

Variable	Variable name	Original US Sample				Original Brazil Sample				Original SA Sample				Original UK Sample			
		Valid		Missing		Valid		Missing		Valid		Missing		Valid		Missing	
		n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
V_1	S1Q1	102	71	41	29	121	61.4	76	39	51	63	30	37	32	41	47	60
V_2	S1Q5	95	66	48	34	108	54.8	89	45	48	59	33	41	31	39	48	61
V_3	S2Q1	102	71	41	29	124	62.9	73	37	52	64	29	36	32	41	47	60
V_4	S2Q5	98	69	45	32	107	54.3	90	46	48	59	33	41	30	38	49	62
V_5	S3Q1	103	72	40	28	117	59.4	80	41	48	59	33	41	32	41	47	60
V_6	S3Q5	96	67	47	33	108	54.8	89	45	47	58	34	42	30	38	49	62
V_7	QS	88	62	55	39	93	47.2	104	53	40	49	41	51	26	33	53	67
V_8	EPQ1	88	62	55	39	93	47.2	104	53	40	49	41	51	26	33	53	67
V_9	EPQ2	87	61	56	39	93	47.2	104	53	40	49	41	51	26	33	53	67
V_{10}	EPQ3	88	62	55	39	93	47.2	104	53	40	49	41	51	26	33	53	67
V_{11}	EPQ5	88	62	55	39	93	47.2	104	53	40	49	41	51	26	33	53	67
V_{12}	EPQ6	88	62	55	39	93	47.2	104	53	40	49	41	51	26	33	53	67

Table G1 cont.

Valid and Missing Data Per Country

Variable	Variable name	Original US Sample				Original Brazil Sample				Original SA Sample				Original UK Sample			
		Valid		Missing		Valid		Missing		Valid		Missing		Valid		Missing	
		n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
V_{13}	EPQ7	88	62	55	39	93	47.2	104	53	40	49	41	51	26	33	53	67
V_{14}	EPQ8	88	62	55	39	89	45.2	108	55	39	48	42	52	26	33	53	67
V_{15}	EPQ9	87	61	56	39	92	46.7	105	53	38	47	43	53	26	33	53	67
V_{16}	EPQ10	87	61	56	39	92	46.7	105	53	40	49	41	51	26	33	53	67
V_{17}	EPQ11	87	61	56	39	92	46.7	105	53	40	49	41	51	26	33	53	67
V_{18}	EPQ12	87	61	56	39	92	46.7	105	53	40	49	41	51	26	33	53	67

Table G2

Valid and Invalid Cases Per Country

	Original US Sample		Original Brazil Sample		Original SA Sample		Original UK Sample	
Completion of variables	*n*	%	*n*	%	*n*	%	*n*	%
[a] 100%	84	58	85	43	36	44	25	32
[b] ≥75%	3	2	8	4	4	5	1	1
[c] <25%	56	39	104	53	41	51	53	67

[a] All variables completed. This is considered a valid case.

[b] More than or equal to 13 out of 18 variables completed. This is considered a valid case.

[c] Less than 5 out of 18 variables completed. This is considered an invalid case.

Appendix H

Coding Scheme

Dimensions	Responses coded to dimension	Code	Synonymous terms coded to dimension
Programme structural features	Clearly specified programme goals	1	Outcomes; purpose; indeterminate programme purpose; well-defined; understandable; changing focus of goals
	Realistic outcomes		Goal alignment; applicable outcomes
	Measurable outcomes		
	Agreement on goals		Consensus on goals; lack of stable/consistent goals; clarity on goals
	Adequate data		Sufficient quality indicator; proper data record
	Reliable data		Consistent data; systematic data
	Accessible data		Available data; lack of data; existing data
	Explicit programme theory		Well explained theory; defined; clear programme theory
	Plausible programme theory		Logic model
	Clearly defined service delivery		
	Clearly defined target beneficiaries		
	Implementation fidelity (programme implemented as planned)		Program structure; variances in implementation; level of implementation

Dimensions	Responses coded to dimension	Code	Synonymous terms coded to dimension
Stakeholder characteristics	Willingness to collaborate	2	Input; resistance; engagement; buy-in; investment; animosity; enthusiasm; support; interest; initiative; commitment; stakeholder readiness; apathy; trust
	Authority to act on findings		Capacity; autonomy
	Transparency about purpose of evaluation		Clarity; language that enables non-experts to understand
Logistical requirements	Adequate budget	3	
	Adequate timeframe		Time; deadline; unrealistic timeframe
	Feasibility of conducting proposed evaluation		Too many programs with different aims and populations; complexity; barriers
	Feasibility to implement desired methodology		Finding comparison group; control data

Note. Words not associated with any themes (e.g., required evaluation) were not coded; Double-barrelled/unclear sub categories with no overriding criteria were not coded; Criteria that did not fall under any of the proposed dimensions were coded

Appendix I

Q Factor Analysis

Descriptive Statistics

Eighty-eight point one percent (88.1%) of the final sample (*n* = 260) completed the Q sort task, resulting in a total of 4351 individual Q sorts (see Table I1 for sample breakdown).

Table I1

Percentage of Participants who Completed the Q Sort Task, and Number of Individual Q Sort per Cohort of Interest

Evaluator Cohort	*n*	% of final sample	No. of individual Q sorts
US	86	91.5	1634
UK	26	86.7	494
Brazil	79	86.8	1501
SA	38	84.4	722
Total	229	88.1%	4351

Inspection of individual Q sorts confirmed that the distribution of Q statements was not concentrated in any particular category for any respondent. In fact, most US (*n* = 83) and SA respondents (*n* = 37), and all UK and Brazil respondents used at least three categories to sort the 19 Q statements. I did not identify any systematic sorting pattern that warranted the exclusion of specific Q sorts for the US, UK, and SA evaluator cohorts. Three problematic cases were however identified among Brazil respondents: one respondent allocated all 19 Q statements to the *Not at all important* category; and two respondents used only the first two categories *Not at all important* and *Quite unimportant* to distribute the Q statements (distribution ratio of 18:1 and 17:2 respectively). These three cases were excluded from subsequent analyses. The internal consistency of the Q sort task was deemed satisfactory for all four cohorts. Cronbach alpha values ranged between .6 and .8.

The *Quite unimportant* category was the most used sorting category, containing 36.9% of the total responses. Participants were able to distinguish between the five different sorting categories and the 19 Q statements (see Table I2).

Table I2

Q Sorts per Sorting Category

	US		*UK*		*Brazil*		*SA*	
Sorting category	*n*	%	*n*	%	*n*	%	*n*	%
Not at all important	481	29.4	137	27.7	639	42.6	236	32.7
Quite unimportant	637	39.0	202	40.9	475	31.6	293	40.6
Neither unimportant nor important	319	19.5	87	17.6	174	11.6	136	18.8
Quite important	120	7.3	43	8.7	93	6.2	33	4.6
Essential	77	4.7	25	5.1	120	8.0	24	3.3

Only 16.5% of the total responses were allocated to the *Neither unimportant nor important* category, with QS12 (programme is implemented as intended) being assigned most frequently to this particular category. QS17 (*timeframe is adequate to complete the evaluation*) was assigned most frequently to the *Not at all important* category by US and SA evaluators, while QS19 (*the required methodology is feasible*) was assigned most frequently to this particular category by UK and Brazil evaluators. Assignment of Q statements to the *Essential* category was more fragmented: QS12 (*programme is implemented as intended*) was assigned more frequently to this particular category by UK and SA evaluators, while QS14 (*stakeholders have authority to act on findings)* and QS9 (*programme theory is plausible) were assigned* more frequently to this category by US and Brazil evaluators, respectively.

Factor Extraction

The initial sorting of the Q statements formed the basis for the Q factor analysis, which was replicated across all four evaluator cohorts. The factorability of the inter-correlation matrix for each cohort was examined. All factored entities had a correlation of at least .3 with multiple other entities, suggesting reasonable factorability (Field, 2013). Communalities were well above .5 (see Table I3 for an example; communalities for the last 10 respondents in the US cohort are presented) further confirming that factored entities shared common variance. The Q factor analysis was therefore performed on the Q sorts of 226 respondents in total (US = 86; UK = 26; Brazil = 76; SA = 38).

Table I3

Communalities for Last ten Respondents in the US Cohort before Rotation

Respondent ID	Initial	Extraction
K_78	1.00	.98
K_79	1.00	.98
K_80	1.00	.98
K_81	1.00	.97
K_82	1.00	.97
K_83	1.00	.97
K_84	1.00	.97
K_85	1.00	1.00
K_86	1.00	.99
K_108	1.00	.97

Between seven and 16 factors with eigenvalues greater than one were extracted (see Tables I4-I5 for examples of initial unrotated PCA solutions).

Table I4

Unrotated Factor Solution for the US Cohort

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	25.61	29.77	29.77	25.61	29.77	29.77
2	11.81	13.73	43.51	11.81	13.73	43.51
3	6.72	7.82	51.32	6.72	7.82	51.32
4	6.26	7.28	58.61	6.26	7.28	58.61
5	5.47	6.35	64.96	5.47	6.35	64.96
6	4.47	5.20	70.16	4.47	5.20	70.16
7	4.12	4.79	74.96	4.12	4.79	74.96
8	3.69	4.29	79.25	3.69	4.29	79.25
9	3.35	3.89	83.14	3.35	3.89	83.14
10	2.67	3.10	86.25	2.67	3.10	86.25
11	2.33	2.71	88.96	2.33	2.71	88.96
12	2.03	2.36	91.31	2.03	2.36	91.31
13	1.64	1.91	93.22	1.64	1.91	93.22
14	1.40	1.62	94.84	1.40	1.62	94.84
15	1.33	1.54	96.39	1.33	1.54	96.39
16	1.24	1.45	97.83	1.24	1.45	97.83
17	25.61	1.13	98.96			
18	11.81	1.04	100.00			

Table I5

Unrotated Factor Solution for the Brazil Cohort

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	17.76	23.37	23.37	17.76	23.37	23.37
2	8.91	11.72	35.09	8.91	11.72	35.09
3	6.76	8.90	43.99	6.76	8.90	43.99
4	6.01	7.91	51.89	6.01	7.91	51.89
5	5.25	6.91	58.80	5.25	6.91	58.80
6	4.64	6.10	64.90	4.64	6.10	64.90
7	4.11	5.41	70.31	4.11	5.41	70.31
8	3.72	4.90	75.20	3.72	4.90	75.20
9	3.14	4.13	79.34	3.14	4.13	79.34
10	3.10	4.07	83.41	3.10	4.07	83.41
11	2.66	3.50	86.90	2.66	3.50	86.90
12	2.12	2.79	89.69	2.12	2.79	89.69
13	1.98	2.61	92.30	1.98	2.61	92.30
14	1.71	2.25	94.55	1.71	2.25	94.55
15	1.42	1.87	96.42	1.42	1.87	96.42
16	1.12	1.47	97.89	1.12	1.47	97.89
17	.95	1.25	99.15			

Only three to six factors were retained for a Varimax rotation (see Table I6 for breakdown of results per evaluator cohort), based on the percentage of variance criterion and scree test criterion. For example, only the first five factors extracted for the US cohort were retained for an orthogonal rotation as they individually explained relatively large amounts of variance (29.8%, 13.7%, 7.8%, 7.2%, and 6.4% respectively) compared to the remaining 11 factors, and cumulatively accounted for 64.9 % of the total variance.

Table I6

Breakdown of Initial Unrotated Q Factor Analysis Results for each Cohort of Interest

Evaluator Cohort	No. factors extracted	No. of factors retained for orthogonal rotation	% of variance explained cumulatively by factors retained for rotation
US	16	5	64.9%
UK	7	3	62.3%
Brazil	16	6	64.9%
SA	10	4	65.5%

Analysis of the rotated factor matrices (see Table I7 for an example; only the first 30 factor loadings for the US cohort are presented) revealed that most respondents' Q sorts loaded highly on only one of the rotated factors, with the first two factors dominating the solution as expected. It should be noted that only factor loadings greater than .40 are displayed in the rotated factor matrixes and retained for interpretation.

Table I7

Initial Rotated Matrix: US Cohort

Respondent	Component				
	1	2	3	4	5
K_74	.91				
K_11	.88				
K_13	.88				
K_80	.86				
K_43	.85				
K_37	.84				
K_31	.82				
K_8	.74				
K_72	.73				
K_34	.69				
K_17	.69				
K_33	.68				
K_86	.67	.47			
K_21	.66				
K_9	.65	.59			
K_59	.65	.43	-.46		
K_22	.63	.47			
K_82	.59				
K_78	.59				
K_41	.58				
K_46	.58			.53	
K_108	.54			-.51	
K_79	.53			-.42	
K_81	.53	.50			
K_45	.52				
K_3	.50		.40		
K_23	.49		.44		
K_2	-.48			-.45	
K_84	.44				
K_36	.43	.42			

Note. Extraction method: Principal component analysis. Rotation method: Varimax with Kaiser normalization.

Problematic cases (e.g., respondents with significant factor loadings not adequately accounted for by the solution) were identified and deleted (see Table I8 for number of problematic cases for each evaluator cohort).

Table I8

Problematic Cases Identified in Initial Rotated Matrix of Each Cohort of Interest

Evaluator Cohort	*n* with factor loadings < .50	*n* with cross loadings	*n* with communalities < .50	No. of times analysis was re-run
US	10	3	6	2
UK	3	2	2	2
Brazil	16	3	2	3
SA	3	3	3	3

The analysis was re-run at least twice until a satisfactory factor solution was obtained. For example, the analysis had to be re-run thrice for the Brazil and SA cohorts to arrive at a factor solution with as many pure factor loadings as possible: the first time after deleting 24 and nine problematic cases respectively; the second time after deleting eight and one problematic case respectively; and the third time after deleting one problematic case from each dataset.

Final Rotated Solutions

Table I9

Final Solution Total Variance Explained: US Cohort

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	20.05	33.41	33.41	14.73	24.55	24.55
2	10.20	17.00	50.42	13.91	23.18	47.72
3	5.50	9.16	59.58	5.37	8.96	56.68
4	4.48	7.46	67.04	5.30	8.83	65.51
5	3.64	6.06	73.10	4.55	7.59	73.10

Note. Extraction method: Principal component analysis

Table I10

Final Solution Total Variance Explained: UK Cohort

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.57	44.51	44.51	7.57	44.51	44.51	5.29	31.09	31.09
2	2.98	17.51	62.03	2.98	17.51	62.03	4.25	25.00	56.09
3	1.92	11.27	73.29	1.92	11.27	73.29	2.92	17.20	73.29
4	.98	5.73	79.03						
5	.79	4.65	83.68						
6	.65	3.83	87.50						
7	.51	2.98	90.48						
8	.47	2.74	93.22						
9	.31	1.80	95.02						

Note. Extraction method: Principal component analysis.

Table I11

Final Solution Total Variance Explained: Brazil Cohort

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	12.62	29.35	29.35	12.62	29.35	29.35	9.60	22.33	22.33
2	6.00	13.94	43.30	6.00	13.94	43.30	6.41	14.92	37.24
3	4.94	11.49	54.78	4.94	11.49	54.78	5.59	13.00	50.24
4	3.68	8.56	63.34	3.68	8.56	63.34	4.47	10.40	60.64
5	3.06	7.11	70.46	3.06	7.11	70.46	3.50	8.13	68.77
6	2.37	5.50	75.96	2.37	5.50	75.96	3.09	7.19	75.96

Note. Extraction method: Principal component analysis.

Table I12

Final Solution Total Variance Explained: SA Cohort

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.87	31.46	31.46	7.87	31.46	31.46	6.96	27.84	27.84
2	4.78	19.11	50.57	4.78	19.11	50.57	4.14	16.55	44.38
3	3.27	13.08	63.65	3.27	13.08	63.65	3.76	15.03	59.42
4	2.36	9.43	73.08	2.36	9.43	73.08	3.42	13.66	73.08
5	1.39	5.55	78.62						
6	1.13	4.53	83.15						

Note. Extraction method: Principal component analysis

Problematic Factors

Table I13

Factor 3 Crib Sheet: US Cohort

Criteria	Q Statements	Factor Scores
Items with the highest rankings in Factor 3 array	Programme theory is plausible (QS9)	2.2
	Programme theory is explicitly stated (QS8)	1.6
Items ranked higher in Factor 3 array than any other factor array	Programme data are easily accessible (QS7)	0.8
	Programme theory is plausible (QS9)	2.2
	Budget is adequate for the evaluation (QS16)	0.2
	Type of evaluation required is feasible(QS18)	0.3
Items ranked lower in Factor 3 array than any other factor array	Programme outcomes are realistic (QS2)	-1.2
	Stakeholders agree on programme goals (QS4)	-1.6
	Target beneficiaries are clearly defined (QS11)	-1.4
	Programme is implemented as intended (QS12)	-0.4
Items with the lowest rankings in Factor 3 array	Stakeholders agree on programme goals (QS4)	-1.6
	Target beneficiaries are clearly defined (QS11)	-1.4
	Programme outcomes are realistic (QS2)	-1.2
	Programme outcomes are measurable (QS3)	-1.0

Table I14

Factor 4 Crib Sheet: US Cohort

Criteria	*Q Statement*	Factor Score
Items with the highest rankings in Factor 4 array	Programme outcomes are measurable (QS3	1.5
	Programme outcomes are realistic (QS2)	1.2
	Stakeholders have authority to act on the findings (QS14)	1.1
Items ranked higher in Factor 4 array than any other factor array	Programme outcomes are realistic (QS2)	1.2
	Programme outcomes are measurable (QS3)	1.5
	Programme data are reliable(QS6)	0.9
	Stakeholders are willing to collaborate with the evaluator (QS13)	0.6
	Timeframe is adequate to complete the evaluation (QS17)	0.3
Items ranked lower in Factor 4 array than any other factor array	The manner in which the programme is delivered is clearly defined (QS 10)	-1.5
	Type of evaluation required (process, outcome or impact) is feasible (QS18)	2.3
Items with the lowest rankings in Factor 4 array	Type of evaluation required (process, outcome or impact) is feasible (QS18)	-2.3
	The manner in which the programme is delivered is clearly defined (QS 10)	-1.5
	Stakeholders agree on program goals (QS4)	-1.0
	Required evaluation methodology is feasible (QS19).	-1.0

Background Characteristics of Evaluators Sharing Common Perspectives

Table I15

Background Characteristics of US Evaluators Sharing Perspectives 1 and 2

	Perspective 1		Perspective 2	
	n	%	*n*	%
Current Involvement in Evaluation				
Design evaluations	19	86.4	21	95.5
Conduct evaluations	19	86.4	22	100.0
Lead team of evaluators	9	40.9	8	36.4
Employed in evaluation job	16	72.6	18	81.8
Academic interest in evaluation	7	31.8	9	40.9
Publish on evaluation	10	45.5	7	31.8
Other	3	13.6	-	-
Employment Setting				
University	2	9.1	6	27.3
Public sector	6	27.3	3	13.6
Private evaluation consultancy firm	3	13.6	5	22.7
Own evaluation consultancy	7	31.8	3	13.6
NGO/NPO	2	9.1	6	27.3
Other	2	9.1	-	-
Highest Academic Qualification				
Undergraduate	1	4.5	-	-
Postgraduate diploma	-	-	-	-
Master's degree	8	36.4	11	50.0
PhD	13	59.1	10	45.5
Other	-	-	1	4.5
Type of Training in Evaluation				
Self-educated	6	27.3	9	40.9
Short course certificate in evaluation	2	9.1	4	18.2
Postgraduate diploma	1	4.5	-	-
Master's degree	5	22.7	5	22.7
PhD	8	36.4	4	18.2
Years Conducting Evaluation				
Less than a year	-	-	-	-
1 to 5 years	6	27.3	11	50.0
6 to 10 years	5	22.7	4	18.2
11 to 15 years	5	22.7	-	
More than 15 years	6	27.3	7	31.8
Level of Experience: Outcome Evaluations				
Slightly experienced	1	4.5	1	4.5
Moderately experienced	9	40.9	5	22.7
Highly experienced	12	54.2	16	72.7

Note. Perspective 1 = Theory-driven (*N* = 22) ; Perspective 2 = Utilisation-focused (*N* = 22)

Table I16

Background Characteristics of UK Evaluators Sharing Perspectives 1 and 2

	Perspective 1		Perspective 2	
	n	%	*n*	%
Current Involvement in Evaluation				
Design evaluations	7	87.5	4	80.0
Conduct evaluations	6	75.0	4	80.0
Lead team of evaluators	6	75.0	3	60.0
Employed in evaluation job	6	75.0	3	60.0
Academic interest in evaluation	4	50.0	-	-
Publish on evaluation	1	12.5	2	40.0
Employment Setting				
University	2	25.0	-	-
Public sector	1	12.5	-	-
Private evaluation consultancy firm	2	25.0	3	60.0
Own evaluation consultancy	2	25.0	1	20.0
NGO/NPO	1	12.5	-	-
Other	-	-	1	20.0
Highest Academic Qualification				
Undergraduate	1	12.5	1	20.0
Postgraduate diploma	-	-	-	-
Master's degree	5	62.5	-	-
PhD	2	25.0	4	80.0
Type of Training in Evaluation				
Self-educated	6	75.0	4	80.0
Short course certificate in evaluation	1	12.5	-	-
Postgraduate diploma	-	-	-	-
Master's degree	-	-	1	20.0
PhD	1	12.5	-	-
Years Conducting Evaluation				
Less than a year	-	-	-	-
1 to 5 years	1	12.5	-	-
6 to 10 years	4	50.0	1	20.0
11 to 15 years	1	12.5	1	20.0
More than 15 years	2	25.0	3	60.0
Level of Experience: Outcome Evaluations				
Slightly experienced	2	25.0	-	-
Moderately experienced	1	12.5	1	20.0
Highly experienced	5	62.5	4	80.0

Note. Perspective 1= Theory-driven (*N* = 8); Perspective 2= Outcome-based and utilisation-focused (*N* = 5).

Table I17

Background Characteristics of Brazil Evaluators Sharing Perspectives 1, 2 and 3

	Perspective 1		Perspective 2		Perspective 3	
	n	%	*n*	%	n	%
Current Involvement in Evaluation						
Design evaluations	6	42.9	3	33.3	4	66.7
Conduct evaluations	10	71.4	4	44.4	3	50.0
Lead team of evaluators	4	28.6	2	22.2	1	16.7
Employed in evaluation job	4	28.6	-	-	1	16.7
Academic interest in evaluation	8	57.1	8	88.9	3	50.0
Publish on evaluation	7	50.0	2	22.7	4	66.7
Other	-	-	1	11.1	-	-
Employment Setting						
University	4	28.6	3	33.3	2	33.3
Public sector	6	42.9	4	44.4	-	-
Private evaluation consultancy	-	-	-	-	1	16.7
Own evaluation consultancy	2	14.3	-	-	1	16.7
NGO/NPO	-	-	2	22.2	1	16.7
Other	2	14.3	-	-	1	16.7
Highest Academic Qualification						
Postgraduate diploma	3	21.4	-	-	1	16.7
Master's degree	5	35.7	4	44.4	4	66.7
PhD	6	42.9	5	55.6	1	16.7
Type of Training in Evaluation						
Self-educated	6	42.9	1	11.1	-	-
Short course certificate in evaluation	4	28.6	5	55.6	4	66.7
Postgraduate diploma	2	14.3	1	11.1	-	-
Master's degree	1	7.1	1	11.1	1	16.7
PhD	1	7.1	1	11.1	1	16.7
Years Conducting Evaluation						
Less than a year	3	21.4	2	22.2	1	16.7
1 to 5 years	6	42.9	3	33.3	1	16.7
6 to 10 years	2	14.3	2	22.2	-	-
11 to 15 years	3	21.4	-	-	2	33.3
More than 15 years	-	-	2	22.2	2	33.3
Level of Experience: Outcome Evaluations						
Slightly experienced	4	28.6	2	22.2	2	33.3
Moderately experienced	6	42.9	5	55.6	2	33.3
Highly experienced	4	28.6	2	22.2	2	33.3

Note. Perspective 1 = Theory-driven (*N* = 14); Perspective 2 = utilisation-focused (*N* = 9); Perspective 3 = Implementation-focused (*N* = 6)

Table I18

Background Characteristics of SA Evaluators Sharing Perspectives 1 and 2

	Perspective 1		Perspective 2	
	n	%	*n*	%
Current Involvement in Evaluation				
Design evaluations	7	70.0	2	33.3
Conduct evaluations	8	80.0	4	66.7
Lead team of evaluators	6	60.0	1	16.7
Employed in evaluation job	6	60.0	2	33.3
Academic interest in evaluation	6	60.0	5	83.3
Publish on evaluation	2	20.0	3	50.0
Employment Setting				
University	-	-	4	66.7
Public sector	1	10.0	-	-
Private evaluation consultancy firm	1	10.0	1	16.7
Own evaluation consultancy	4	40.0	-	-
NGO/NPO	4	40.0	1	16.7
Highest Academic Qualification				
Undergraduate	1	10.0	-	-
Postgraduate diploma	-	-	-	-
Master's degree	6	60.0	2	33.3
PhD	3	30.0	4	66.7
Type of Training in Evaluation				
Self-educated	4	40.0	4	66.7
Short course certificate in evaluation	1	10.0	1	16.7
Postgraduate diploma	1	10.0	1	16.7
Master's degree	3	30.0	-	-
PhD	1	10.0	-	-
Years Conducting Evaluation				
Less than a year	1	10.0	-	-
1 to 5 years	3	30.0	2	33.3
6 to 10 years	3	30.0	1	16.7
11 to 15 years	2	20.0	1	16.7
More than 15 years	1	10.0	2	33.3
Level of Experience: Outcome Evaluations				
Slightly experienced	2	20.0	1	20.0
Moderately experienced	4	40.0	2	40.0
Highly experienced	4	40.0	2	40.0

Note. Perspective 1 = Theory-driven (*N* = 10); Perspective 2 = utilisation-focused (*N* = 6)

Appendix J

Correspondence Analysis

Two-way Contingency Tables

Table J1

Correspondence Table: US Cohort

	Study Tasks				
Evaluability Criterion	Scenario 1	Scenario 2	Scenario 3	QSort	Total
QS1	**7**	**7**	**12**	**13**	**39**
QS2	0	1	1	4	6
QS3	0	0	0	5	5
QS4	2	8	8	1	19
QS5	1	7	2	5	15
QS6	1	1	1	2	5
QS7	16	6	6	1	29
QS8	1	0	0	3	4
QS9	3	2	3	0	8
QS10	0	0	0	1	1
QS11	0	0	0	0	0
QS12	0	0	1	1	2
QS13	**20**	**11**	**25**	**12**	**68**
QS14	1	3	0	2	6
QS15	1	0	0	5	6
QS16	1	1	1	6	9
QS17	**27**	**29**	**12**	**10**	**78**
QS18	0	0	0	10	10
QS19	2	3	2	1	8
Total	83	79	74	82	318

Note. QS1= Programme goals are clearly specified; QS2=Programme outcomes are realistic; QS3= Programme outcomes are measurable; QS4= Stakeholders agree on programme goals; QS5= Programme data are adequate; QS6= Programme data are reliable; QS7= Programme data are easily accessible; QS8= Programme theory is explicitly stated; QS9= Programme theory is plausible; QS10= The manner in which the programme is delivered is clearly defined; QS11= Target beneficiaries are clearly defined; QS12= Programme is implemented as intended; QS13= Stakeholders are willing to collaborate with the evaluator; QS14= Stakeholders have authority to act on evaluation findings;QS15= Stakeholders are transparent about the purpose of the evaluation; QS16= Budget is adequate for the evaluation; QS17= Timeframe is adequate to complete the evaluation; QS18= Type of evaluation required is feasible; QS19= Required evaluation methodology is feasible.

Table J2

Correspondence Table: UK Cohort

	Study Tasks				
Evaluability Criterion	Scenario 1	Scenario 2	Scenario 3	QSort	Total
QS1	**5**	**4**	**5**	**4**	**18**
QS2	0	0	0	1	1
QS3	1	0	0	2	3
QS4	0	3	1	0	4
QS5	0	1	0	1	2
QS6	0	1	0	1	2
QS7	3	2	1	1	7
QS8	0	0	0	0	0
QS9	0	0	1	0	1
QS10	0	0	0	0	0
QS11	0	0	0	0	0
QS12	0	0	0	0	0
QS13	**7**	**2**	**7**	**1**	**17**
QS14	2	1	1	2	6
QS15	0	0	0	1	1
QS16	0	1	1	5	7
QS17	**1**	**8**	**4**	**0**	**13**
QS18	0	0	0	4	4
QS19	1	1	1	2	5
Total	20	24	22	25	91

Note. QS1= Programme goals are clearly specified; QS2=Programme outcomes are realistic; QS3= Programme outcomes are measurable; QS4= Stakeholders agree on programme goals; QS5= Programme data are adequate; QS6= Programme data are reliable; QS7= Programme data are easily accessible; QS8= Programme theory is explicitly stated; QS9= Programme theory is plausible; QS10= The manner in which the programme is delivered is clearly defined; QS11= Target beneficiaries are clearly defined; QS12= Programme is implemented as intended; QS13= Stakeholders are willing to collaborate with the evaluator; QS14= Stakeholders have authority to act on evaluation findings;QS15= Stakeholders are transparent about the purpose of the evaluation; QS16= Budget is adequate for the evaluation; QS17= Timeframe is adequate to complete the evaluation; QS18= Type of evaluation required is feasible; QS19= Required evaluation methodology is feasible.

Table J3

Correspondence Table: Brazil Cohort

Evaluability Criterion	Study Tasks				Total
	Scenario 1	Scenario 2	Scenario 3	QSort	
QS1	**16**	**14**	**18**	**26**	**74**
QS2	0	0	1	0	1
QS3	0	0	2	6	8
QS4	1	3	4	4	12
QS5	1	4	2	3	10
QS6	0	0	0	4	4
QS7	10	2	1	0	13
QS8	1	0	0	1	2
QS9	2	1	1	1	5
QS10	1	0	0	2	3
QS11	0	0	0	3	3
QS12	0	0	0	1	1
QS13	**16**	**14**	**9**	**5**	**44**
QS14	2	2	0	2	6
QS15	0	0	0	5	5
QS16	3	1	2	1	7
QS17	**6**	**6**	**5**	**3**	**20**
QS18	0	0	0	5	5
QS19	1	1	1	4	7
Total	60	48	46	76	230

Note. QS1= Programme goals are clearly specified; QS2=Programme outcomes are realistic; QS3= Programme outcomes are measurable; QS4= Stakeholders agree on programme goals; QS5= Programme data are adequate; QS6= Programme data are reliable; QS7= Programme data are easily accessible; QS8= Programme theory is explicitly stated; QS9= Programme theory is plausible; QS10= The manner in which the programme is delivered is clearly defined; QS11= Target beneficiaries are clearly defined; QS12= Programme is implemented as intended; QS13= Stakeholders are willing to collaborate with the evaluator; QS14= Stakeholders have authority to act on evaluation findings;QS15= Stakeholders are transparent about the purpose of the evaluation; QS16= Budget is adequate for the evaluation; QS17= Timeframe is adequate to complete the evaluation; QS18= Type of evaluation required is feasible; QS19= Required evaluation methodology is feasible.

Table J4

Correspondence Table: SA Cohort

Evaluability Criterion	Study Tasks				Total
	Scenario 1	Scenario 2	Scenario 3	QSort	
QS1	**6**	**8**	**7**	**7**	**28**
QS2	0	0	0	1	1
QS3	1	0	0	4	5
QS4	3	3	4	2	12
QS5	1	0	1	1	3
QS6	0	0	1	1	2
QS7	5	3	3	0	11
QS8	2	2	3	4	11
QS9	1	0	2	1	4
QS10	0	0	0	0	0
QS11	0	0	0	0	0
QS12	1	0	0	0	1
QS13	**5**	**7**	**5**	**5**	**22**
QS14	0	1	0	0	1
QS15	0	0	0	0	0
QS16	1	0	1	2	4
QS17	**8**	**8**	**4**	**3**	**23**
QS18	0	0	0	3	3
QS19	2	4	3	1	10
Total	36	36	34	35	141

Note. QS1= Programme goals are clearly specified; QS2=Programme outcomes are realistic; QS3= Programme outcomes are measurable; QS4= Stakeholders agree on programme goals; QS5= Programme data are adequate; QS6= Programme data are reliable; QS7= Programme data are easily accessible; QS8= Programme theory is explicitly stated; QS9= Programme theory is plausible; QS10= The manner in which the programme is delivered is clearly defined; QS11= Target beneficiaries are clearly defined; QS12= Programme is implemented as intended; QS13= Stakeholders are willing to collaborate with the evaluator; QS14= Stakeholders have authority to act on evaluation findings;QS15= Stakeholders are transparent about the purpose of the evaluation; QS16= Budget is adequate for the evaluation; QS17= Timeframe is adequate to complete the evaluation; QS18= Type of evaluation required is feasible; QS19= Required evaluation methodology is feasible.

Correspondence Maps: US Cohort

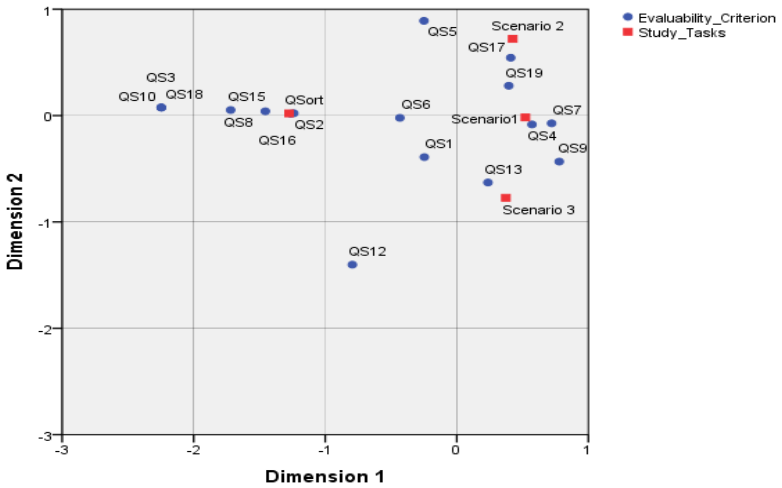


Figure J1. CA map with all data points

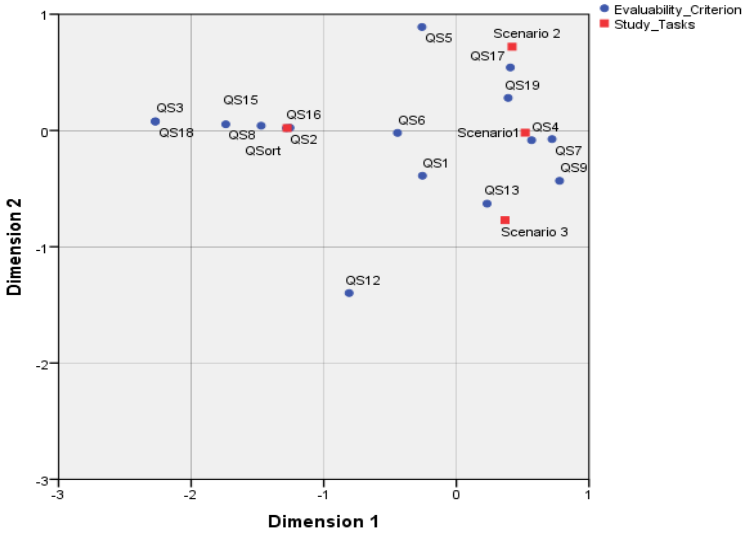


Figure J2. CA map excluding QS10 (low frequency point)

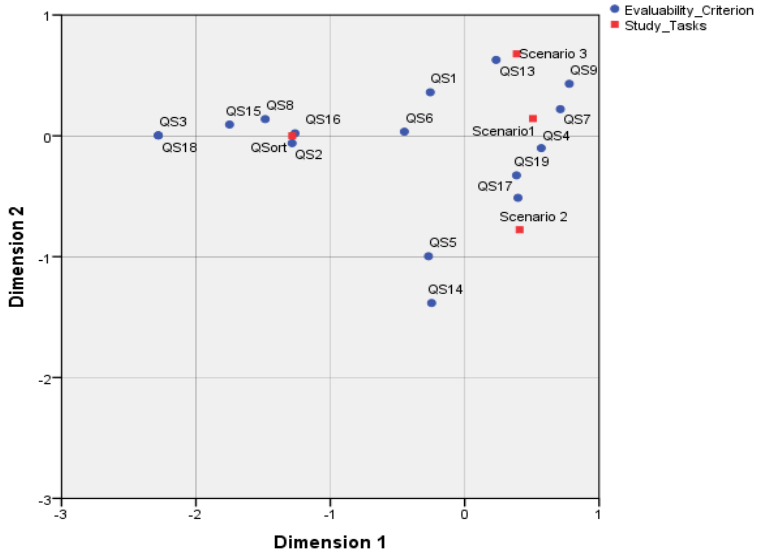


Figure J3. CA map excluding low frequency point and QS12 (outlier)

Row and Column Coordinates for Two-dimensional Solution: US Cohort

Table J4

Row Coordinates: US Correspondence Map

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
Evaluability Criterion	Mass	1	2	Inertia	1	2	1	2	Total
QS1	.124	-.256	.361	.012	.014	.061	.382	.354	.736
QS2	.019	-1.285	-.061	.019	.056	.000	.947	.001	.948
QS3	.016	-2.280	.003	.047	.147	.000	.996	.000	.996
QS4	.060	.570	-.101	.032	.035	.002	.349	.005	.354
QS5	.048	-.270	-.997	.019	.006	.179	.102	.655	.757
QS6	.016	-.449	.034	.002	.006	.000	.996	.003	.999
QS7	.092	.712	.221	.047	.083	.017	.553	.025	.578
QS8	.013	-1.485	.138	.018	.050	.001	.854	.003	.858
QS9	.025	.778	.431	.010	.027	.018	.873	.125	.999
QS11	.000	.	.	.	.	.	.	.	.
QS13	.216	.233	.629	.030	.021	.324	.221	.752	.973
QS14	.019	-.247	-1.384	.010	.002	.138	.064	.935	.998
QS15	.019	-1.750	.093	.035	.104	.001	.937	.001	.938
QS16	.029	-1.263	.020	.026	.081	.000	.998	.000	.998
QS17	.248	.396	-.513	.043	.069	.247	.510	.402	.911
QS18	.032	-2.280	.003	.093	.293	.000	.996	.000	.996
QS19	.025	.385	-.327	.003	.007	.010	.645	.218	.863
Active Total	1.000			.446	1.000	1.000			

Table J5

Column Coordinates: US Cohort

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
Study Tasks	Mass	1	2	Inertia	1	2	1	2	Total
Scenario1	.263	.508	.143	.075	.121	.020	.509	.019	.527
Scenario 2	.251	.408	-.777	.069	.074	.575	.339	.575	.914
Scenario 3	.232	.387	.678	.066	.062	.404	.298	.428	.725
QSort	.254	-1.283	.001	.236	.743	.000	.999	.000	.999
Active Total	1.000			.446	1.000	1.000			

Correspondence Maps for Analysis: UK Cohort

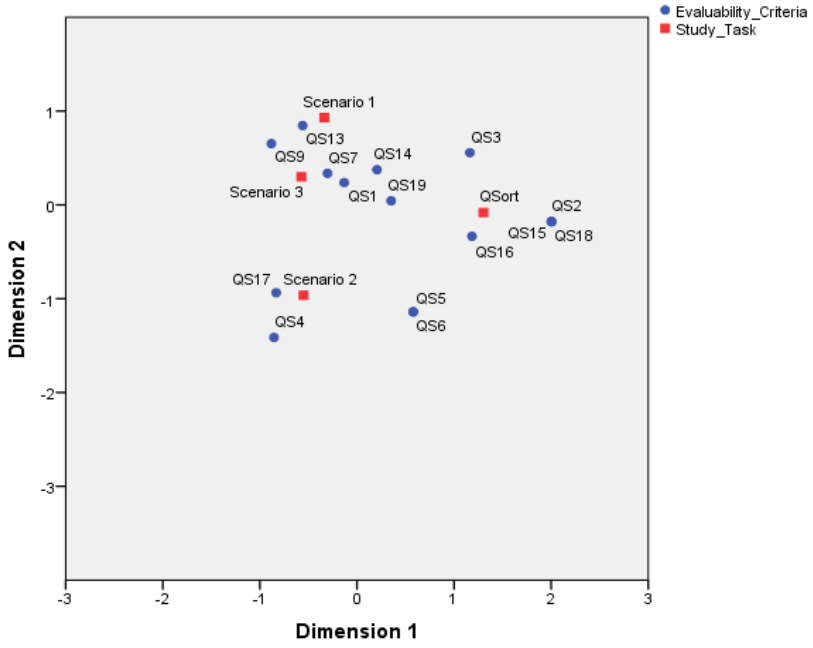


Figure J4. CA map with all data points

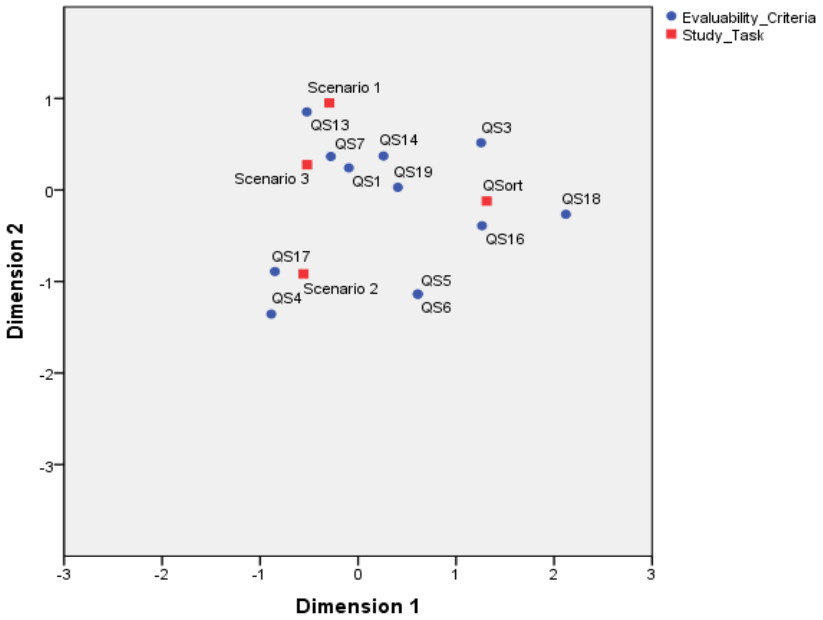


Figure J5. CA *map excluding QS2, QS9, and QS15 (low frequency points).*

Correspondence Maps for the Brazil Cohort

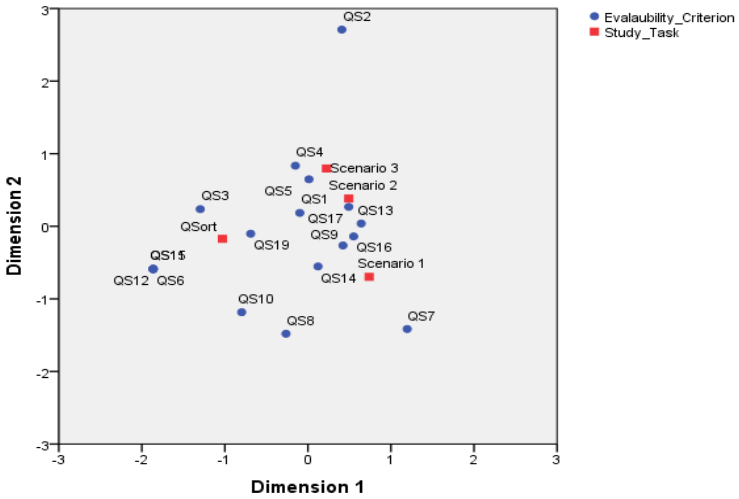


Figure J6. CA map with all data points

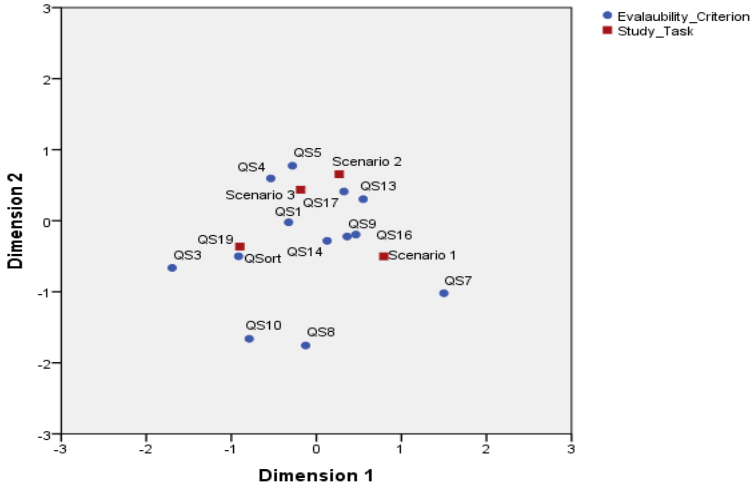


*Figure J7. C*A map excluding QS2, QS6, QS11, QS12, QS15 and QS18 (low frequency points)

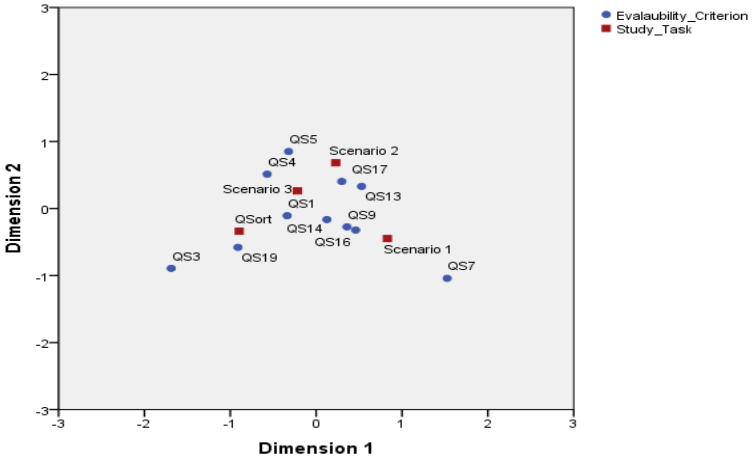


*Figure J8. C*A map excluding low frequency points, and QS10 and QS8 (outliers)

Correspondence Maps for the SA Cohort

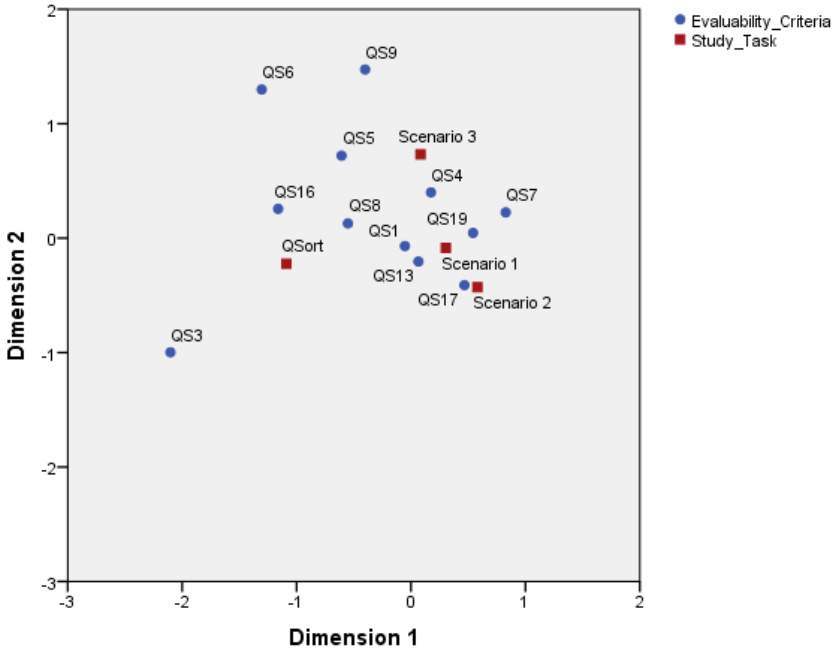


Figure J9. CA Map excluding QS2, QS12, QS14 and QS18 (low frequency points)

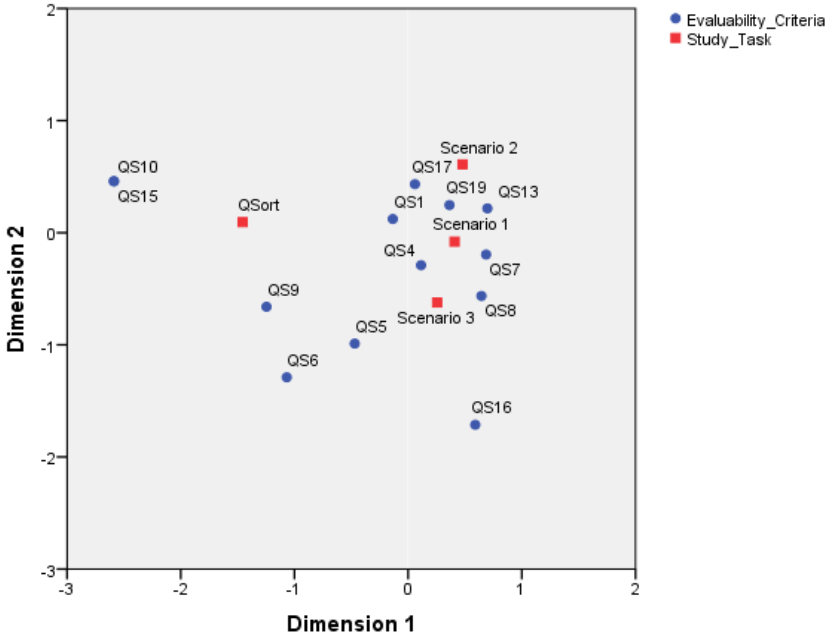


Figure J10. CA Map excluding low frequency points and QS3 (Outlier)

Appendix K

Multinomial Logistic Regression

Goodness of Fit and Pseudo R^2 Estimates for D for Model 1 (Assessment of Evaluability)

Table K1

Goodness of Fit for Assessment of Evaluability

	χ^2	*df*	*p*
Scenario 1			
Pearson	12.118	8	.146
Deviance	11.363	8	.182
Scenario 2			
Pearson	7.932	8	.440
Deviance	8.181	8	.416
Scenario 3			
Pearson	6.240	8	.620
Deviance	6.210	8	.624

Table K2

Pseudo R Square Assessment of Evaluability

	Scenario 1	Scenario 2	Scenario 3
Cox and Snell	.105	.045	.033
Nagelkerke	.119	.056	.047

Observed and Predicted Frequencies (Assessment of Evaluability)

Table K3

Classification Table for Scenario 1 (DV: Assessment of Evaluability)

	Predicted			
Observed	Low	Medium	High	Percent Correct
Low	43.000	2.000	32.000	55.8%
Medium	22.000	3.000	22.000	6.4%
High	27.000	4.000	72.000	69.9%
Overall Percentage	40.5%	4.0%	55.5%	52.0%

Table K4

Classification Table for Scenario 2 (DV: Assessment of Evaluability)

	Predicted			
Observed	Low	Medium	High	Percent Correct
Low	0	0	30.000	0.0%
Medium	0	0	37.000	0.0%
High	0	0	158.000	100.0%
Overall Percentage	0.0%	0.0%	99.9%	70.2%

Table K5

Classification Table for Scenario 3 (DV: Assessment of Evaluability)

	Predicted			
Observed	Low	Medium	High	Percent Correct
Low	0	0	11.000	0.0%
Medium	0	0	34.000	0.0%
High	0	0	178.000	100.0%
Overall Percentage	0.0%	0.0%	99.9%	79.7%

Calculation of Proportional by Chance Accurate Rate for Model 1 (Assessment of Evaluability)

The proportional by chance accuracy rate for each scenario was calculated by adding up the squared marginal percentages of the dependent variable. The marginal percentages are presented in Tables K6-K8.

Table K6

Marginal Percentages for Scenario 1 (DV: Assessment of Evaluability)

		N	Marginal Percentage
Scenario 1_Evaluability level	Low	77	33.9%
	Medium	47	20.7%
	High	103	45.4%
Experience (in years)	Low	100	44.1%
	Medium	51	22.5%
	High	76	33.5%
Practice context	In developing countries	122	53.7%
	In developed countries	90	39.6%
	In both developed and developing countries	15	6.6%
Valid		227	100.0%
Missing		33	
Total		260	
Subpopulation		9	

Table K7

Marginal Percentages for Scenario 2 (DV: Assessment of Evaluability)

		N	Marginal Percentage
Scenario 2_Evaluability level	Low	30	13.3%
	Medium	37	16.4%
	High	158	70.2%
Experience (in years)	Low	98	43.6%
	Medium	51	22.7%
	High	76	33.8%
Practice context	In developing countries	121	53.8%
	In developed countries	90	40.0%
	In both developed and developing countries	14	6.2%
Valid		225	100.0%
Missing		35	
Total		260	
Subpopulation		9	

Table K8

Marginal Percentages for Scenario 3 (DV: Assessment of Evaluability)

		N	Marginal Percentage
Scenario 3_Evaluability level	Low	11	4.9%
	Medium	34	15.2%
	High	178	79.8%
Experience (in years)	Low	98	43.9%
	Medium	51	22.9%
	High	74	33.2%
Practice context	In developing countries	121	54.3%
	In developed countries	88	39.5%
	In both developed and developing countries	14	6.3%
Valid		223	100.0%
Missing		37	
Total		260	
Subpopulation		9	

The proportional by chance accuracy rate was 0.36 ($0.339^2 + 0.207^2 + 0.454^2$) for scenario 1, 0.54 ($0.133^2 + 0.164^2 + 0.702^2$) for scenario 2, and 0.65 ($0.049^2 + 0.152^2 + 0.789^2$) for scenario 3. The proportional by chance criteria was 45% (1.25 x 0.36) for scenario 1; 67.5% (1.25 x 0.54) for scenario 2, and 81.3% (1.25 x 0.65) for scenario 3. Overall, the model accurately predicted 52%, 70.2%, and 79.7% of the cases for Scenario 1, Scenario 2, and Scenario 3 respectively. The *High difficulty* level category had the highest level of accurate

prediction, ranging between 69.9% and 100%, compared to the other two categories. This model was therefore most useful for predicting this particular category. This finding is not surprising as MLR tends to produce the most accurate predictions for the largest categories (Petrucci, 2009). The classification accuracy rates for Scenario 1 and Scenario 2 were above the proportional by chance accuracy criteria, suggesting the model was useful for predicting the cases in these two scenarios. The classification accuracy rate for Scenario 3 was below the proportional by chance accuracy criteria, suggesting that the overall criterion of classification accuracy was not satisfied (the explanatory variables do not contribute significantly to the explanation of the dependent variable in Scenario 3).

Goodness of Fit and Pseudo R^2 Estimates for D for Model 2 (Likelihood of Conducting Evaluation)

Table K9

Goodness of Fit for Likelihood of Conducting Evaluation

	χ^2	*df*	*p*
Scenario 1			
Pearson	5.177	8	.739
Deviance	5.353	8	.719
Scenario 2			
Pearson	5.664	8	.685
Deviance	5.745	8	.676
Scenario 3			
Pearson	3.477	8	.901
Deviance	3.617	8	.890

Table K10

Pseudo R Square for Likelihood of Conducting Evaluation

Scenario 1	Scenario 2	Scenario 3
.128	.039	.002
.145	.046	.002

Parameter Estimates for Scenario 2 (Likelihood of Evaluating Programme)

Table K11

Parameter Estimates (DV: Likelihood of Evaluating Programme)

Likelihood of evaluating program (Scenario 2)[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Low likelihood	Intercept	1.028	.723	2.020	1	.155			
	Low Experience	-.214	.426	.252	1	.616	.807	.350	1.862
	Medium Experience	.267	.532	.252	1	.616	1.307	.460	3.710
	High Experience	0[b]	.	.	0	.	.	.	.
	Developing Context	-.260	.715	.132	1	.716	.771	.190	3.132
	Developed Context	1.086	.776	1.960	1	.162	2.963	.648	13.555
	Both	0[b]	.	.	0	.	.	.	.
Moderate Likelihood	Intercept	.442	.801	.305	1	.581			
	Low Experience	-.183	.501	.134	1	.714	.832	.312	2.220
	Medium Experience	.255	.613	.173	1	.677	1.290	.388	4.288
	High Experience	0[b]	.	.	0	.	.	.	.
	Developing Context	-.535	.794	.453	1	.501	.586	.124	2.779
	Developed Context	.323	.862	.141	1	.707	1.382	.255	7.483
	Both	0[b]	.	.	0	.	.	.	.

[a]The reference category is: High Likelihood.
[b]This parameter is set to zero because it is redundant.

Observed and Predicted Frequencies (Likelihood of Evaluating Programme)

Table K12

Classification Table for Scenario 1 (DV: Likelihood of Evaluating Programme)

	Predicted			
Observed	Low likelihood	Moderate Likelihood	High Likelihood	Percent Correct
Low likelihood	70.000	7.000	14.000	76.9%
Moderate Likelihood	20.000	14.000	21.000	25.5%
High Likelihood	36.000	9.000	36.000	44.4%
Overall %	55.3%	13.2%	31.1%	52.6%

Table K13

Classification Table for Scenario 1 (DV: Likelihood of Evaluating Programme)

	Predicted			
Observed	Low likelihood	Moderate Likelihood	High Likelihood	Percent Correct
Low likelihood	136.000	0	0	100.0%
Moderate Likelihood	48.000	0	0	0.0%
High Likelihood	41.000	0	0	0.0%
Overall %	99.3%	0.0%	0.0%	60.0%

Table K14

Classification Table for Scenario 1 (DV: Likelihood of Evaluating Programme)

	Predicted			
Observed	Low likelihood	Moderate Likelihood	High Likelihood	Percent Correct
Low likelihood	144.000	0	0	100.0%
Moderate Likelihood	51.000	0	0	0.0%
High Likelihood	28.000	0	0	0.0%
Overall %	99.3%	0.0%	0.0%	64.1%

Calculation of Proportional by Chance Accurate Rate for Model 2 (Likelihood of Evaluating Programme)

The proportional by chance accuracy rate was 0.35 ($0.401^2 + 0.242^2 + 0.357^2$) for scenario 1, 0.44 ($0.604^2 + 0.213^2 + 0.182^2$) for scenario 2, and 0.49 ($0.646^2 + 0.229^2 + 0.126^2$) for scenario 3. The proportional by chance criteria was 43.8% (1.25 x 0.35) for scenario 1; 55% (1.25 x 0.44) for scenario 2, and 61.3% (1.25 x 0.49) for scenario 3 (see Tables K15-K17 below). The classification accuracy rate for all three scenarios were above the proportional by chance criteria, suggesting that the overall criterion of classification accuracy was satisfied (the explanatory variables contribute significantly to the explanation of the dependent variable for all three scenarios).

Table K15

Marginal Percentages for Scenario 1 (DV: Likelihood of Evaluating Programme)

		N	Marginal Percentage
Scenario 1_ Likelihood of Evaluating Programme	Low	91	40.1%
	Medium	55	24.2%
	High	81	35.7%
Experience (in years)	Low	100	44.1%
	Medium	51	22.5%
	High	76	33.5%
Practice context	In developing countries	122	53.7%
	In developed countries	90	39.6%
	In both developed and developing countries	15	6.6%
Valid		227	100.0%
Missing		33	
Total		260	
Subpopulation		9	

Table K16

Marginal Percentages for Scenario 2 (DV: Likelihood of Evaluating Programme)

		N	Marginal Percentage
Scenario 2_ Likelihood of Evaluating Programme	Low	136	60.4%
	Medium	48	21.3%
	High	41	18.2%
Experience (in years)	Low	98	43.6%
	Medium	51	22.7%
	High	76	33.8%
Practice context	In developing countries	121	53.8%
	In developed countries	90	40.0%
	In both developed and developing countries	14	6.2%
Valid		225	100.0%
Missing		35	
Total		260	
Subpopulation		9	

Table K17

Marginal Percentages for Scenario 1 (DV: Likelihood of Evaluating Programme)

		N	Marginal Percentage
Scenario 3_ Likelihood of Evaluating Programme	Low	144	64.6%
	Medium	51	22.9%
	High	28	12.6%
Experience (in years)	Low	98	43.9%
	Medium	51	22.9%
	High	74	33.2%
Practice context	In developing countries	121	54.3%
	In developed countries	88	39.5%
	In both developed and developing countries	14	6.3%
Valid		223	100.0%
Missing		37	
Total		260	
Subpopulation		9	